

**Ein Beitrag zum rechnergestützten
Entwurf rekursiver digitaler Filter
minimalen Aufwands**

von

Günter Friedrich Dehner

Erlangen, 1976

1976 № 373

Ein Beitrag zur rechnergestützten
Entwurf rekursiver digitaler Filter
mit Hilfe von Algorithmen

Digitized by the Internet Archive
in 2026

Ein Beitrag zum rechnergestützten Entwurf rekursiver digitaler Filter minimalen Aufwands

Dem Fachbereich Ingenieurwissenschaften
der Universität Erlangen-Nürnberg
zur Erlangung des Grades
Doktor - Ingenieur

vorgelegt von

Günter Friedrich Dehner

Erlangen, 1976



Als Dissertation genehmigt von
Fachbereich Ingenieurwissenschaften
der Universität Erlangen - Nürnberg

Tag der Einreichung: 12. März 1976
Tag der Promotion : 14. Juli 1976

Dekan: Prof.Dr.rer.nat. H.J.Schneider
Berichterstatter: Prof.Dr.-Ing. H.W.Schüßler
Prof.Dr.phil.nat. H.Schlitt

Herrn Prof. Dr.-Ing. H.W. Schüßler gebührt mein Dank für die Anregung zur Arbeit auf dem Gebiet der Digitalen Filter und die sich hieraus ergebende interessante Themenstellung, für seine Hilfe bei der Suche nach Lösungswegen, für zahlreiche Hinweise und Diskussionen, sowohl während der Arbeit als auch bei der Zusammenstellung der Ergebnisse.

Herrn Prof. Dr. phil.nat. H. Schlitt danke ich für sein Interesse an dieser Arbeit.

INHALTSVERZEICHNIS

	Seite
VERZEICHNIS DER WICHTIGSTEN FORMELZEICHEN	5
1. EINFÜHRUNG	7
2. GRUNDLAGEN ZUM ENTWURF REKURSIVER DIGITALER FILTER	10
2.1 Beschreibung eines digitalen Systems	10
2.2 Verschiedene Approximationsverfahren	16
2.2.1 Lösung der Approximationsaufgabe durch Optimierung	16
2.2.2 Geschlossene Lösungen für den Entwurf selektiver digitaler Filter	19
2.3 Unterschiedliche Darstellung und Quantisierung der Koeffizienten	26
2.3.1 Allgemeine Zahlendarstellung der Koeffizienten	26
2.3.2 Bemerkungen zur Realisierung der verschiedenen Koeffizienten-Darstellungen	32
3. ENTWURF SELEKTIVER DIGITALER FILTER MINIMALEN KOEFFIZIENTENAUFWANDS BEI KASKADENSCHALTUNG	34
3.1 Untersuchung der Koeffizientenempfindlichkeit im Durchlaß- und Sperrbereich	35
3.1.1 Bestimmung des maximalen Fehlers in Abhängigkeit von der Koeffizientenwortlänge	35
3.1.2 Abhängigkeit des maximalen Fehlers von der Entwurfskonstanten C	39
3.1.3 Abschätzung einer theoretischen Fehlergrenze	42
3.2 Optimierung der notwendigen Koeffizientenwortlänge durch Variation des Entwurfsparameters C	48
3.2.1 Adaptive Suchoptimierung	49
3.2.2 Programmsystem zum optimierten Filterentwurf	53
3.2.3 Ausblick auf weitere Entwurfsmöglichkeiten	53
3.3 Minimierung des Koeffizientenaufwands	54
3.3.1 Schaltungstechnische Bewertung der Koeffizientenwortlänge	54

3.3.2 Die Auswirkung der Filtergrader- weiterung auf den Realisierungsaufwand	58
3.4 Beispiele zur Koeffizientenoptimierung	59
3.4.1 Cauer-Tiefpässe mit unterschiedlicher Filtergraderweiterung	59
3.4.2 Tschebyscheff-Tiefpässe mit unter- schiedlicher Filtergraderweiterung	61
3.4.3 Verschiedene Approximationen der Band- sperrung nach Golden und Kaiser	63
3.4.4 Tiefpaß 10. Grades mit Skalierung	66
4. THEORETISCHE GRUNDLAGEN FÜR DIE BEWERTUNG DES RAUSCHVERHALTENS VERSCHIEDENER STRUKTUREN MIT UNTERSCHIEDLICHER ARITHMETIK	67
4.1 Zahlendarstellung der Zustandsvariablen	68
4.2 Modelle zur Berechnung des Quantisierungs- rauschens bei unterschiedlichen Rundungs- arithmetiken	69
4.2.1 Modelle mit kontinuierlicher Fehlerver- teilung	70
4.2.1.1 Runden	72
4.2.1.2 Zweierkomplement-Abschneiden	74
4.2.1.3 Betragsmäßiges Abschneiden	76
4.2.1.4 Kontrolliertes Runden	79
4.2.2 Modelle mit diskreter Fehlerverteilung	82
4.2.2.1 Zweierkomplement-Runden	83
4.2.2.2 Betragsmäßiges Runden	86
4.2.2.3 Mathematisches Runden	88
4.2.2.4 Zweierkomplement-Abschneiden	88
4.2.2.5 Unsymmetrisches betragsmäßiges Abschneiden	90
4.2.2.6 Symmetrisches betragsmäßiges Abschneiden	91
4.2.3 Berechnung der Rauschleistung eines Systems mit mehreren Nichtlinearitäten	93
4.2.3.1 Rauschartiger Fehleranteil	96
4.2.3.2 Mit dem Signal korrelierter Fehleran- teil	96
4.2.3.3 Gleichanteil des Fehlers	102

4.3 Dynamikbereich eines digitalen Systems	102
4.3.1 Notwendige Prüfstellen	103
4.3.2 Aussteuerungskriterien	104
4.4 Rauschzahl und zusätzliche innere Wortlänge Δw	109
5. ANALYTISCHER VERGLEICH DES RAUSCHVERHALTENS	
UNTERSCHIEDLICHER REALISIERUNGEN	112
5.1 Übersicht über die untersuchten Strukturen	112
5.1.1 Einführung	112
5.1.2 Katalog der untersuchten Strukturen	113
5.2 Untersuchung der Grundblöcke 2. Grades	122
5.2.1 Beschreibung der durchgeführten Tests	124
5.2.2 Rauschverhalten rein rekursiver Strukturen	125
5.2.3 Das Rauschverhalten unterschiedlicher Strukturen bei Runden	126
5.2.3.1 Katalog der Ergebnisse	128
5.2.3.2 Vergleich der Ergebnisse	141
5.2.4 Rauschverhalten bei unterschiedlicher Arithmetik	144
5.2.4.1 Runden	145
5.2.4.2 Zweierkomplement-Abschneiden	148
5.2.4.3 Betragsmäßiges Abschneiden	150
5.2.5 Rauschverhalten grenzyklenfreier Strukturen	153
5.2.6 Das Rauschverhalten von Strukturen mit rückgekoppelten Speichern und inneren Skalierern	159
5.2.7 Das Rauschverhalten bei 2er-Potenz-Skalierung und anschließender Verkürzung	164
5.3 Untersuchung und Optimierung der Kaskadenschaltung	169
5.3.1 Verteilung und Verteilungsdichte der Aufwandsmaße bei vollständiger Enumeration aller möglichen Anordnungen der Zähler- und Nennerblöcke	169
5.3.2 Feste Zuordnung der Pol- und Nullstellen	177
5.3.3 Optimierung der Zuordnung und Reihenfolge der Teilblöcke	179
5.3.3.1 Dynamisches Programmieren	180
5.3.3.2 Optimierung mit begrenztem Zwischenspeicher	182

5.3.3.3 Vergleich der Ergebnisse verschiedener Verfahren	185
5.3.3.4 Abschätzung der Rechenzeiten	188
5.3.3.5 Einschränkungen bei der Optimierung von Strukturen mit korrelierten Fehleranteilen	193
5.3.4 Vergleich der Optimierungsergebnisse bei unterschiedlichen Realisierungen der Teilblöcke	197
5.3.4.1 Verschiedene Realisierungen der 1. und 2. kanonischen Form bei Runden	198
5.3.4.2 Vergleich weiterer Realisierungen	202
5.3.5 Einfluß der Entwurfsparameter auf das Rauschverhalten einer Kaskade	204
5.3.5.1 Einfluß der Filtergraderweiterung	204
5.3.5.2 Abhängigkeit von der Entwurfskonstanten C	206
6. SIMULATION DIGITALER SYSTEME	208
6.1 Programm zur Simulation digitaler Systeme	208
6.2 Strukturuntersuchung mittels Simulation	210
6.2.1 Bestimmung der Aussteuerung	212
6.2.2 Bestimmung der Gesamtrauschleistung	213
6.2.3 Bestimmung der mit dem Signal linear unkorrelierten Rauschleistung	215
6.2.4 Notwendige Simulationslänge	216
6.3 Ergebnisse verschiedener Strukturen	218
6.3.1 Grundstrukturen 2. Grades	218
6.3.2 Das Rauschverhalten von Tiefpässen 8. und 10. Grades	224
7. ZUSAMMENFASSUNG	228
8. LITERATURVERZEICHNIS	230
ANHANG: Datensätze der Entwurfsbeispiele	A 1
1. Tiefpässe	A 1
2. Bandpässe	A 6
3. Bandsperren	A 7

VERZEICHNIS DER WICHTIGSTEN FORMELZEICHEN

a_v	Koeffizient, allgemein
A	Gesamtaufwand
A_C, A_N, A_S, A_R	versch. Teilaufwandsfaktoren
b_v, d_v, g_v, h_v	Zählerkoeffizienten
$B(z)$	Zählerübertragungsfunktion
c_v, e_v	Nennerkoeffizienten
C, c	Entwurfskonstante, normierte -
$C(z)$	Nennerübertragungsfunktion
e	Exponent eines "Pseudo"-Gleitkomma-Koeffizienten
$e^{(=)}, e^{(s)}$	Kennwerte für Gleichanteil des Fehlers, -- mit dem Signal korrelierten Fehleranteil
f, f_A	Frequenzvariable, Abtastfrequenz
$F(z), v_{F_i}(z)$	Übertragungsfunktionen vom Eingang zu den Übersteuerungsstellen bzw. zu den Nichtlinearitäten
$G(z), v_{G_i}(z)$	Fehlerübertragungsfunktionen
$h_o(k)$	Impulsantwort
$H(z), v_{H_\lambda}(z)$	Übertragungsfunktionen
$H(\Omega)$	Abkürzung für $H(e^{j\Omega})$
$H_W(\Omega)$	gewünschtes Übertragungsverhalten
$H_W(\Omega)$	Übertragungsfunktion mit gerundeten Koeffizienten
i_M, i_E	Mantissen-Länge, Exponenten -
K	Verstärkungsfaktor
$K(\Omega)$	charakteristische Funktion
l	Anzahl der Blöcke 2. Grades
m_R	Anzahl der Multiplizierer
n	Filtergrad
N	Fehlerleistung
$N^{(R)}, N^{(S)}, N^{(=)}$	versch. Teilfehlerleistungen
p	ld des Zahlenbereichs
$p(x), P(x)$	Verteilungsdichte, Verteilungsfunktion
$P(\Omega)$	Leistungsübertragungsfunktion
q, q_C, q_S	Länge der Koeffizienten bzw. Zustandsvariablen nach dem Festkomma
Q, Q_C, Q_S	Quantisierung

R, R_j	Rauschzahl, Teilrauschzahl
R_i, R_{iO}	innere Rauschzahl, - des linear unkorrelierten Fehlers
$S_{\max}$	maximale Zwischenspeichergröße
T_v	abgeschätzte Rechenzeit
$u(k)$	Eingangssignal
w, w_c, w_s	Wortlänge, - der Koeffizienten, - der Zustandsvariablen
$x(k)$	Zustandsvariable
$y(k)$	Ausgangssignal
z	komplexe Frequenzvariable digitaler Systeme
Z	Zahlenbereich
$\gamma_O(k)$	Impuls
δ_D, δ_S	Fehlerstrahlen des Toleranzschemas
Δ_1, Δ_2	Fehlerstrahlen der charakt. Funktion
$\Delta(k), \Delta_i(k)$	Quantisierungsfehler
$\Delta^{(R)}, \Delta^{(S)}, \Delta^{(=)}$	versch. Anteile des Quantisierungsfehlers
$\overline{\Delta^2}$	mittleres Fehlerquadrat
Δn	Filtergraderweiterung
Δw	zusätzliche innere Wortlänge
$\varepsilon_D(\Omega), \varepsilon_S(\Omega)$	Fehlerübertragungsfunktionen
$\varepsilon_{Dw}, \varepsilon_{Sw}$	Fehler der Übertragungsfunktion bei verschiedener Koeffizientenwortlänge
ρ_N, ρ_P	Nullstellen-, Pol-Radius
$\phi_{ij}(k)$	Auto-, bzw. Kreuzkorrelationsfunktion
$\phi_{ij}(z)$	spektrale Leistungsdichte
$\psi(\Omega), \psi_i(\Omega)$	Winkelübertragungsfunktion, - von Eingang bis zur Nichtlinearität
ψ_N, ψ_P	Nullstellen-, Pol-Winkel
σ^2, σ	Varianz, Streuung
$\sigma^2(R)$	Varianz des rauschartigen Fehleranteils
Ω	normierte Frequenz
Ω_D, Ω_S	normierte Frequenzgrenzen.

1. Einführung

Der Einsatz digitaler Systeme, speziell digitaler Filter wird heute auf den verschiedensten Arbeitsgebieten der Signalverarbeitung diskutiert. Für die endgültige Verwirklichung der Vorhaben ist unter anderem eine ökonomische - d.h. kostensparende - Realisierung der digitalen Systeme Voraussetzung.

Eine Bewertung des Realisierungsaufwandes hängt dabei grundsätzlich von der Art der Realisierung ab. So wird der Entwurf eines Filters zur Implementierung auf einem Universalrechner unter anderen Gesichtspunkten erfolgen, als der für einen speziellen Filteraufbau.

Die endliche Genauigkeit der Rechenoperationen - endliche Länge der Koeffizienten und Zustandsvariablen - führt beim Betrieb digitaler Filter zu Abweichungen vom theoretischen Verhalten.

Im Gegensatz zu analogen Schaltungen können diese Fehler durch hinreichende Wortlängen der Koeffizienten und Zustandsvariablen beliebig klein gehalten werden. Bei vorgegebenem Universalrechner bedeutet die Verlängerung der Arithmetik eine größere Rechenzeit und bei der Realisierung eines speziellen Filters mit vorgegebener Arbeitsgeschwindigkeit eine umfangreichere Schaltung - in beiden Fällen also einen höheren Realisierungsaufwand.

Aufbauend auf der Anzahl der mathematischen Verknüpfungen und der Länge der Koeffizienten und Zustandsvariablen ist ein Aufwandsmaß zu finden, das zur Beurteilung der Filterentwürfe verwendet werden kann.

Eine möglichst genaue Analyse des Fehlerverhaltens digitaler Systeme ist Voraussetzung für die Angabe dieses Aufwandsmaßes und eine Optimierung der Filterentwürfe.

Speziell werden in der vorliegenden Arbeit rekursive digitale Filter untersucht, deren grundlegende Systemeigenschaften in [1] - [5] beschrieben sind.

Für die Realisierung einer vorgegebenen Übertragungsfunktion liegen heute eine Vielzahl verschiedener Strukturen vor. Koeffizientenempfindlichkeit und Rundungsrauschen hängen im allgemeinen von der Struktur ab. Die Darstellung der Koeffizienten hat auch bei sonst gleichem linearen Signalflußgraphen einen Einfluß auf deren Fehlerverhalten. Verschiedene Darstellungsmöglichkeiten werden im Abschnitt 2 anhand einer verallgemeinernden "Pseudo"-Gleitkomma-Darstellung diskutiert.

In Ergänzung zu den Untersuchungen von Avenhaus [6] wird im Abschnitt 3 die Abhängigkeit der Koeffizientenfehler von den Parametern eines analytischen Entwurfsverfahrens geprüft. Durch die Variation eines Parameters wird eine Minimisierung des Koeffizienten-Aufwandes möglich.

Die Verwendung verschiedener Rundungsarithmetiken und deren Lage innerhalb der Strukturen haben weiter einen Einfluß auf die Betriebsfehler der gewählten Realisierungen.

Neben den von Eckhardt in [7] untersuchten Arithmetiken werden im Abschnitt 4 neue Rundungsvorschriften und deren Fehlerverhalten bei kurzen Koeffizienten - bzw.

Zweier-Potenz-Multiplikationen - vorgestellt. Unter Berücksichtigung des Dynamikbereichs wird auch hier ein geeignetes schaltungstechnisches Bewertungsmaß eingeführt.

Die Realisierung höhergradiger Systeme kann in einer Kaskadenanordnung von Teilfiltern maximal 2. Grades erfolgen. Die Untersuchung verschiedener Teilfilterstrukturen im Abschnitt 5 liefert eine erste Selektion der verschiedenen Realisierungsmöglichkeiten.

Bekannterweise hängt das Rauschverhalten einer Kaskadenschaltung von der Zuordnung der Pol- und Nullstellen zu Blöcken 2. Grades und deren Reihenfolge ab. Durch eine systematische Untersuchung aller möglichen Anordnungen werden Kriterien für die Bewertung bekannter und die Entwicklung eines neuen in Bezug auf die Rechenzeit sehr effektiven Optimierungsverfahrens gewonnen.

Optimierungsprogramme zum rechnergestützten Entwurf werden beschrieben. Anhand eines geeigneten Gesamtaufwandsmaßes werden die verschiedenen Approximationen und Realisierungen miteinander verglichen.

Zur praktischen Erprobung der Filter wird abschließend ein Programm zur Simulation digitaler Systeme vorgestellt, das die Nachbildung der verschiedenen Strukturen und Rundungsarithmetiken ermöglicht. Am Beispiel der Rauschleistungsmessung wird die gute Übereinstimmung der analytischen Ergebnisse mit dem tatsächlichen Verhalten gezeigt.

1. Grundlagen zum Entwurf rekursiver digitaler Filter

Wie bereits einleitend erwähnt, liegt heute neben einer Vielzahl von Aufsätzen über verschiedene Teilgebiete eine nahezu abgeschlossene Theorie diskreter Systeme in Buchform vor (z.B. [1] - [5]). Auf eine ausführliche Darstellung der allgemeinen theoretischen Grundlagen digitaler Systeme wird daher im Rahmen der vorliegenden Arbeit verzichtet. Nur die wichtigsten Begriffe und Gleichungen, die für die weiteren Abhandlungen häufiger benötigt werden, sollen kurz im nächsten Abschnitt zusammengeestellt werden.

Anschließend werden die verschiedenen Approximationsmethoden zur Bestimmung einer in vorgegebenen Toleranzgrenzen liegenden Übertragungsfunktion erläutert und deren Eignung für verschiedene Aufgaben diskutiert. Speziell wird auf den Entwurf von Filtern mit Buttersworth-Verhalten, Tschebyscheff-Verhalten im Durchlaß oder Sperrbereich, sowie Cauer-Filtern näher eingegangen.

Wird der Entwurf des Filters mit einer, nur durch den verwendeten Rechner bedingten Wortlänge durchgeführt, so stellt sich bei der Realisierung des Filters die Frage nach der notwendigen Koeffizienten-Wortlänge für die Erfüllung gestellter Forderungen, wenn der damit verbundene schaltungstechnische Realisierungsaufwand minimiert werden soll. Dabei führen verschiedene Darstellungen der Koeffizienten zu unterschiedlichen Ergebnissen.

1.1 Beschreiben eines digitalen Systems

Bei Übertragung einer Information in einem digitalen System werden der kontinuierlichen Zeitfunktion $u(t)$ zu äquidistanten Zeitpunkten $k \cdot T$ Amplitudenproben entnommen, die eine Folge von digital codierten Artstärkenwerten $\{u_k\}$ bilden.

$$u(k) \hat{=} u(k \cdot T) \quad \text{für} \quad k = 0(1)\infty, \quad (2.1)$$

$$u(k) = 0 \quad \text{für} \quad k < 0$$

Die Betrachtung diskreter, digital arbeitender Systeme, im speziellen "digitale Filter" genannt, bleibt im Rahmen der vorliegenden Arbeit auf reelle, kausale und zeitinvariante Systeme beschränkt [1].

Das diskrete System bestimmt dann aus $\{u(k)\}$ die Ausgangsfolge $\{y(k)\}$.

$$y(k) \hat{=} y(kT) \quad k = 0(1)\infty \quad (2.2)$$

In Bezug auf ihre Übertragungseigenschaften sind die zu entwerfenden Systeme als linear zu betrachten. Die aufgrund der endlichen Zahlendarstellung auftretenden nichtlinearen Quantisierungsfehler werden später durch Modelle beschrieben, die an den verschiedensten Stellen des Systems voneinander und vom Signal abhängige sowie unabhängige Fehler einspeisen, die ebenfalls eine lineare Bewertung erfahren. Der Überlagerungssatz bleibt in diesem Modell also sowohl für Signal, als auch für die Fehler des Systems erhalten.

Unter diesen Voraussetzungen sind die hier betrachteten digitalen Systeme durch lineare Differenzengleichungen mit reellen, konstanten Koeffizienten zu beschreiben. Allgemein erhält man für ein rekursives System vom Grad n

$$y(k) = \sum_{\mu=0}^m b_{\mu} u(k+\mu-n) - \sum_{v=0}^{n-1} c_v y(k+v-n), \quad (2.3)$$

dabei ist aus Kausalitätsgründen stets $m \leq n$ zu wählen.

Das Übertragungssystem nach Gl. (2.3) sei zunächst "in Ruhe". Erregt man es hierauf mit einem Impuls

$$u(k) = \gamma_0(k) = \begin{cases} 1 & \text{für } k = 0 \\ 0 & \text{für } k \neq 0 \end{cases} \quad (2.4)$$

so erhält man die Impulsantwort $h_0(k)$

$$y(k) = h_0(k) = \sum_{\mu=0}^m b_{\mu} \gamma_0(k+\mu-n) - \sum_{v=0}^{n-1} c_v y(k+v-n). \quad (2.5)$$

Durch Z-Transformation der Impulsantwort $h_0(k)$, (z.B. nach [5]), erhält man die Übertragungsfunktion des digitalen Systems

$$H(z) = \frac{\sum_{\mu=0}^m b_{\mu} z^{\mu}}{\sum_{v=0}^n c_v z^v} \quad (2.6)$$

mit $c_n = 1$.

Ein Übertragungssystem ist stabil, wenn mit Gl. (2.5) gilt

$$\sum_{k=0}^{\infty} |h_0(k)| < M < \infty \quad (2.7)$$

Diese Bedingung ist erfüllt, wenn die Laurent-Reihenentwicklung von $H(z)$ für $|z| \geq 1$ konvergiert, d.h. alle Pole der Übertragungsfunktion $H(z)$ nach Gl. (2.6) im Inneren des Einheitskreises der z -Ebene liegen.

Eine mögliche Realisierung der Übertragungsfunktion nach Gl. (2.6) zeigt der Signalflußgraph für die "direkte Form" im Bild 2.1a für den Fall $m = n$.

Diese Struktur wird im weiteren mit 1. kanonischer Form bezeichnet. Durch Transponierung des Signalflußgraphen findet man die hierzu transponierte Struktur, die 2. kanonische Form (Bild 2.1b).

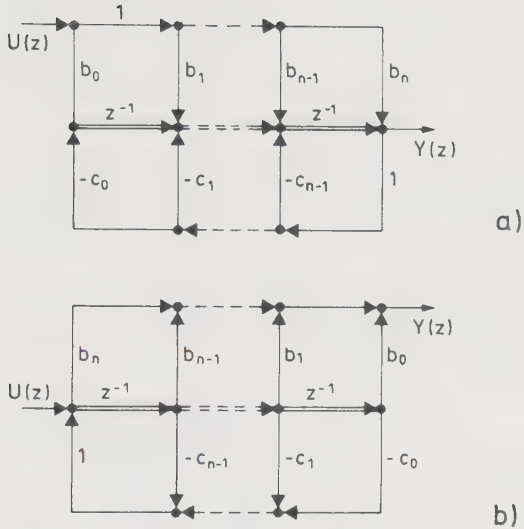


Bild 2.1: Erste (a) und zweite (b) kanonische Form eines diskreten Systems.

Werden die reellen oder konjugiert komplexen Polstellen eines gegebenen Übertragungssystems mit $z_{\infty v}$ und die Nullstellen mit $z_{0\mu}$ bezeichnet, so kann die Übertragungsfunktion $H(z)$ nach Gl. (2.6) als Funktion der Pol- und Nullstellen dargestellt werden.

$$H(z) = b_m \cdot \frac{\prod_{\mu=1}^m (z - z_{0\mu})}{\prod_{v=1}^n (z - z_{\infty v})} \quad (2.8)$$

Konjugiert komplexe Pol- oder Nullstellen lassen sich zu Blöcken 2. Grades mit reellen Koeffizienten zusammenfassen. Mit $\ell = \left[\frac{n+1}{2} \right]^1$ und $m = n$ erhält man

¹⁾ $[x]$ bezeichnet die größte ganze Zahl $\leq x$

$$H(z) = b_m \cdot \prod_{\lambda=1}^{\ell} \frac{z^2 + b_{1\lambda} z + b_{0\lambda}}{z^2 + c_{1\lambda} z + c_{0\lambda}} \quad (2.9)$$

Diese Darstellung entspricht der unskalierten Kaskadenschaltung von Blöcken 2. Grades in der 1. kanonischen Form (Bild 2.2a) und in der 2. kanonischen Form (Bild 2.2b). Bei ungeradem Filtergrad $n = 2\ell - 1$ erhält man für einen Block $b_{0\ell} = c_{0\ell} = 0$ und somit eine Reduktion auf einen Block ersten Grades.

Führt man eine Partialbruchentwicklung von $H(z)$ durch, so erhält man unter der Voraussetzung, daß alle Pole verschieden sind nach [1]

$$H(z) = R_{\infty} + \sum_{v=1}^n \frac{R_v}{z - z_{\infty v}}$$

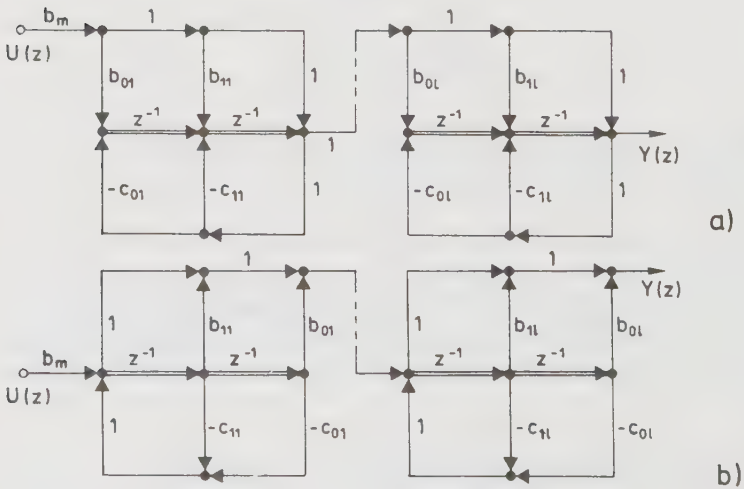


Bild 2.2: Kaskadenschaltung ohne Skalierung

- a) Grundblöcke in der 1. kanonischen Form
- b) Grundblöcke in der 2. kanonischen Form

Dabei lassen sich wiederum konjugiert komplexe Polpaare zu Systemen 2. Grades $H_{\lambda}(z)$ zusammenfassen. Mit $R_{\infty} = b_n$ erhält man

$$H(z) = b_n + \sum_{\lambda=1}^{\ell} H_{\lambda}(z) \quad (2.10)$$

Eine Anwendung findet die letztgenannte Darstellung bei der Realisierung der Übertragungsfunktion in der sogenannten Parallelschaltung (Bild 2.3).

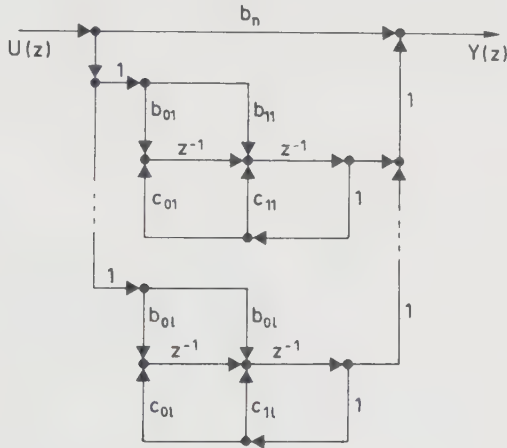


Bild 2.3: Parallelschaltung ohne Skalierung

Nach [1] erhält man den Frequenzgang eines digitalen Übertragungssystems als das Verhältnis der komplexen Amplituden des Eingangs- und Ausgangssignals bei einer Erregung $z = e^{j\omega T}$ im eingeschwungenen Zustand. Mit

$$z = e^{j\omega T} = e^{j\Omega} \quad ; \quad |z| = 1$$

erhält man

$$H(\Omega) \hat{=} H(e^{j\Omega}) = \frac{\sum_{\mu=0}^m b_{\mu} e^{j\mu\Omega}}{\sum_{\nu=0}^n c_{\nu} e^{j\nu\Omega}} = |H(\Omega)| \cdot e^{-j\Psi(\Omega)} \quad (2.11)$$

$$0 \leq \Omega \leq 2\pi$$

den Frequenzgang des digitalen Systems. $\Omega = \omega T$ wird als die normierte Frequenz bezeichnet.

Für den Entwurf minimalphasiger selektiver rekursiver digitaler Filter, mit dem sich die folgenden Abschnitte speziell befassen, ist der Betrag des Frequenzgangs $|H(\Omega)|$ von besonderer Bedeutung. Für den Entwurf von Allpässen oder impulsformenden Netzwerken sind darüber hinaus der Winkel $\Psi(\Omega) = -\angle H(\Omega)$ und die Gruppenlaufzeit $\tau_g(\Omega) = \frac{d\Psi}{d\Omega}$ sowie das Zeitverhalten $h_o(k)$ von Interesse.

2.2 Verschiedene Approximationsverfahren

Der ein Filter beschreibende Koeffizientensatz ist so zu bestimmen, daß die an das Übertragungsverhalten gestellten Anforderungen mit geringstem Aufwand erfüllt werden. Dabei hängt die Bewertung vom Grad des Filters und der Art des jeweils tolerierten Fehlers (Fehlerkriterium) sowie möglichen Nebenbedingungen ab.

Wird das Verhalten des zu entwerfenden Übertragungssystems durch eine Wunschfunktion $H_w(\Omega)$ im Frequenzbereich oder $h_w(k)$ im Zeitbereich angegeben, so können die verschiedenen aus der Literatur bekannten Verfahren [8] - [34] zur Approximation verwendet werden. Sie liefern im allgemeinen bei verschiedenen Fehlerkriterien unterschiedliche Lösungen.

2.2.1 Lösung der Approximationsaufgabe durch Optimierung

Anforderungen an den Betrag der Übertragungsfunktion bzw. an den Dämpfungsgang eines digitalen Filters werden häufig in Form eines Toleranzschemas vorgegeben. Dabei ist es im Prinzip gleichgültig, ob die Schranken bereichsweise konstant sind oder als Kurvenstücke vorliegen, wie im Beispiel von Bild 2.4. Neben dieser Darstellung ist auch eine punktweise Vorgabe der Amplitudenwerte mit zulässigen Toleranzgrenzen für diskrete Frequenzwerte denkbar.

Zur Approximation derart willkürlicher Toleranzschemata eignen sich Optimierungsverfahren nach dem "minimum p-Fehler"-Kriterium, wie sie in [8] - [9] angegeben werden. In [9] wird von Deczky ein Programm zur Optimierung des Frequenzgangs unter der Nebenbedingung angegeben, daß in einem Teilbereich des Toleranzschemas eine "equal ripple" Lösung angestrebt wird. In weiteren Arbeiten [11] - [14] wird die "lineare Programmierung" benutzt. Beispiele und zusätzliche Literaturhinweise sind diesen Arbeiten zu entnehmen.

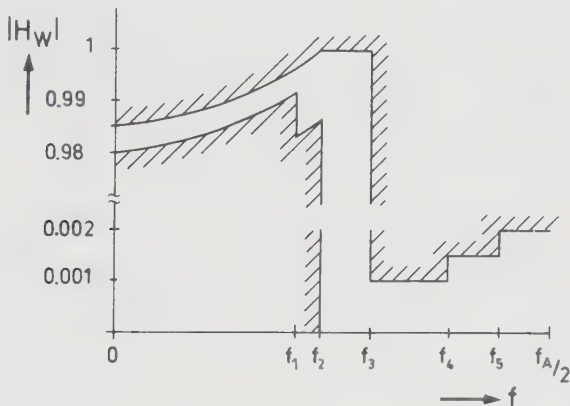


Bild 2.4: Beispiel eines beliebigen Toleranzschemas

Neben dem Dämpfungsgang eines Filters kann auch das Gruppenlaufzeitverhalten von Interesse sein. Entsprechende Beispiele finden sich in [8].

Wird von dem zu entwerfenden Filter ein bestimmtes Zeitverhalten verlangt (ist z.B. die Impulsantwort vorgeschrieben), so führt eine Approximation im Zeitbereich zu einem geeigneten Koeffizientensatz. Je nach Formulierung des Problems kann aus einer begrenzten Anzahl von Werten der Impulsantwort eine direkte Lösung errechnet werden [15] oder mittels eines Optimierungsverfahrens (z.B. nach dem mittleren Fehlerquadrat) eine Approximation der Impulsantwort vorgenommen werden. Weitere Verfahren über die Approximation des Zeitverhaltens mittels

Optimierung sind [16] - [17] zu entnehmen.

Alle vorgenannten Entwurfsverfahren sowie die noch zu behandelnden auf der Approximation eines entsprechenden analogen Filters beruhenden Verfahren gehen zunächst von einer im Prinzip unbegrenzten Koeffizientenwortlänge aus, die meist im Rechner hinreichend genau erreicht werden kann. Bei der Realisierung des digitalen Filters wird man sich jedoch aus Aufwandsgründen auf eine endliche und möglichst geringe Koeffizientenwortlänge [1] beschränken. Durch die nachträgliche Quantisierung der Koeffizienten entstehen zusätzliche Fehler, die den vorher bestimmten Frequenzgang oder das Zeitverhalten mehr oder minder stark verändern können. Dabei hängt der Koeffizientenfehler sowohl von der verwendeten Struktur [1], [6] als auch von der Darstellung der Koeffizienten ab (siehe Abschnitt 2.3).

In den Arbeiten [18] - [19] wird daher durch entsprechende Suchoptimierungsverfahren (z.B. "random-search") der begrenzten Koeffizientenwortlänge bereits bei der Approximation Rechnung getragen. Die Anwendung verschiedener Verfahren der nichtlinearen diskreten Optimierung wird in den Arbeiten [20], [21] vorgeschlagen.

Bei den direkt im diskreten Parameterraum arbeitenden Entwurfsverfahren wirkt sich die erhebliche Rechenzeit, besonders bei größeren Filtersystemen nachteilig aus. Häufig hängt die Anzahl der notwendigen Iterationsschritte auch von der Wahl einer geeigneten Anfangslösung ab. Verwendet man das Ergebnis einer Approximation mit unbegrenzten Koeffizienten als Ausgangslösung, so werden in [6], [22]-[25] unterschiedliche Verfahren zur Nachoptimierung im diskreten Parameterraum vorgeschlagen. In den zitierten Arbeiten wird stets so optimiert, daß die Dämpfung das gegebene Toleranzschema einhält. Durch entsprechende Umformulierung der Fehlerfunktion können diese Verfahren auch auf eine Optimierung der Phase oder der Gruppenlaufzeit, sowie des Zeitverhaltens erweitert werden.

Werden speziell für eine Approximation des Dämpfungsgangs bereichsweise konstante Wunschfunktionen gefordert (z.B. selektiver Tief-, Hoch- und Bandpaß sowie Bandsperre), so erscheint

es sinnvoll, zur Bestimmung der Anfangslösung auf die klassischen analytischen Entwurfsverfahren analoger Filter zurückzugreifen [26]. Ein Vergleich derartiger Verfahren mit dem Entwurf durch reine Optimierung und eine Zusammenstellung der wichtigsten Literaturstellen werden von Sablatash in [27] gegeben.

2.2.2 Geschlossene Lösungen für den Entwurf selektiver digitaler Filter

Der Entwurf selektiver digitaler Filter greift auf die bekannten Verfahren der analogen Filtertechnik zurück [28] - [31]. Dabei erscheint es prinzipiell gleichgültig, ob man die für analoge Filter gefundenen, in der w -Ebene arbeitenden Verfahren in den z -Bereich überträgt [32] oder mittels der bilinearen Transformation die für ein digitales System gestellte Aufgabe in die für ein analoges System überführt und die gefundenen Lösungen anschließend zurücktransformiert [1], [2], [33].

Der Entwurf von Hochpässen sowie symmetrischen Bandpässen und Bandsperren kann durch Allpaßtransformation im z -Bereich oder durch entsprechende Reaktanztransformationen im w -Bereich auf den Entwurf eines normierten Tiefpasses zurückgeführt werden. Zur Approximation unsymmetrischer Bandpässe und Bandsperren kann die Zdunek-Transformation verwendet werden [33] - [35]. Ein Spezialfall der Zdunek-Transformation wird in dem im Abschnitt 3.2 beschriebenen Entwurfsprogramm zur Filtergradverdopplung verwendet. Damit lassen sich auch höhergradige Cauerfilter berechnen, deren Parameter u.U. bei gegebener Rechnergenauigkeit im direkten Verfahren nicht bestimmt werden können. In [36] wird dieses Problem durch eine iterative Berechnung (Newton-Verfahren) der Koeffizienten ohne die Verwendung von elliptischen Funktionen umgangen.

In dem in Abschnitt 3.2 beschriebenen Optimierungsprogramm für digitale Filter minimaler Koeffizientenwortlänge werden die un-

gerundeten Koeffizienten nach dem zweiten Verfahren - Approximation des entsprechenden analogen Filters - bestimmt [1], [26]. Neben der Approximation von Cauer-Filtern [37] kann der Entwurf von Filtern mit Potenz-Verhalten sowie Tschebyscheff-Verhalten im Durchlaß- und Sperrbereich durchgeführt werden. Der Übersicht halber werden die Verfahren am Beispiel eines Tiefpasses kurz erläutert.

Sind die Anforderungen an das Filter durch Angabe der tolerierten Abweichungen von der Wunschfunktion im Durchlaßbereich und im Sperrbereich sowie der Frequenzgrenzen f_D und f_S gegeben, so kann nach Wahl einer geeigneten Abtastfrequenz [1], [33] das Toleranzschema in das auf 2π normierte überführt werden (siehe Bild 2.5a).

Der Betrag der gesuchten normierten Übertragungsfunktion kann als eine Beziehung der charakteristischen Funktion $K(e^{j\Omega})$: = $K(\Omega)$ ausgedrückt werden.

$$|H(e^{j\Omega})|^2 = |H(\Omega)|^2 = \frac{1}{1+C^2|K(\Omega)|^2} \quad (2.12)$$

Je nach gewählter Approximationsart müssen nun die charakteristische Funktion und die Konstante C so bestimmt werden, daß das Toleranzschema von $|H(\Omega)|$ eingehalten wird [1], [2]. Das Toleranzschema der Funktion $K(\Omega)$ zeigt Bild 2.5b.

Die Toleranzschranken δ_D, δ_S gehen dabei in Δ_1 und Δ_2 über.

$$\Delta_1 = \frac{\sqrt{2\delta_D - \delta_D^2}}{1 - \delta_D} \quad (2.13a)$$

$$\Delta_2 = \frac{\sqrt{1 - \delta_S^2}}{\delta_S} \quad (2.13b)$$

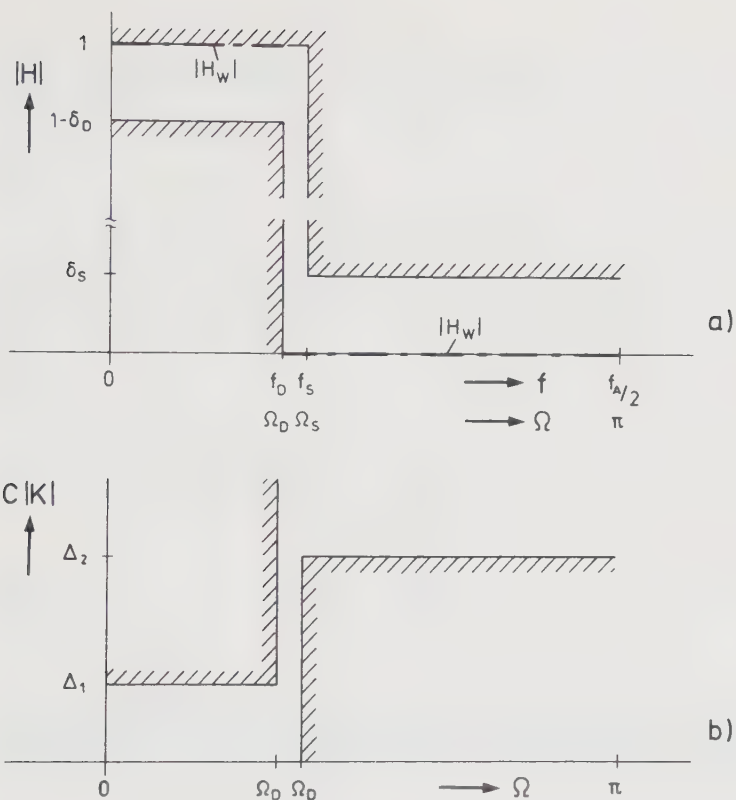


Bild 2.5: a) Übertragungsfunktion $|H(\Omega)|$, $|H(f)|$
b) charakteristische Funktion $C|K(\Omega)|$

Toleranzschema eines Tiefpasses mit

$$f_D = 3,33 \text{ kHz}, \quad f_S = 3,76 \text{ kHz}, \quad f_A = 20 \text{ kHz}$$

$$\Omega_D = \pi/3, \quad \Omega_S = 1,18$$

$$\delta_D = 0,02, \quad \delta_S = 0,001$$

Die charakteristische Funktion ist nun so zu bestimmen, daß stets gilt:

$$C|K| \leq \Delta_1 \quad \text{für} \quad 0 \leq |\Omega| \leq \Omega_D, \quad (2.14a)$$

$$C|K| \geq \Delta_2 \quad \text{für} \quad \Omega_S \leq |\Omega| \leq \pi. \quad (2.14b)$$

Gleichzeitig ergibt sich aus diesen Bedingungen der für die verschiedenen Approximationsarten notwendige Filtergrad $n_{\min}$.

$$n_{\min} = n_{\min}(\Delta_1, \Delta_2, \Omega_D, \Omega_S), \in \mathbb{R} \quad (2.15)$$

Da der Filtergrad stets ganzzahlig ist, muß mindestens die nächstgrößere ganze Zahl gewählt werden. Wie in Abschnitt 3 gezeigt wird, kann die Wahl eines höheren Filtergrades als erforderlichlich sinnvoll sein.

$$n \geq n_{\min} \quad (2.16)$$

Transformiert man das Toleranzschema nach Bild 2.5 weiter in den normierten w -Bereich mit der Durchlaßgrenze $v_D = 1$, so erhält man für die normierte Sperrgrenze

$$v_S = \frac{\tan \frac{\Omega_S}{2}}{\tan \frac{\Omega_D}{2}} \quad (2.17)$$

Die notwendigen Filtergrade $n_{\min}$ für unterschiedliche Approximationen sind den entsprechenden Gleichungen in Tabelle 2.1 zu entnehmen.

Durch die ganzzahlige Wahl des Filtergrades wird das Toleranzschema nach Bild 2.5a nicht voll ausgenutzt. Je nach Vorgabe der Konstanten C innerhalb der Grenzen $C_{\min}$ und $C_{\max}$ kann dabei eine stärkere oder schwächere Ausnutzung des Durchlaß- oder Sperrbereichs erzielt werden.

$$C_{\min} \leq C \leq C_{\max} \quad (2.18)$$

Speziell bedeutet $C = C_{\min}$ eine vollständige Ausnutzung des Toleranzschemas im Sperrbereich und $C = C_{\max}$ eine im Durchlaßbereich. Die Grenzen bei verschiedener Approximation sind ebenfalls Tab. 2.1 zu entnehmen. Eine günstige Auswahl der Entwurfskonstanten wird in Abschnitt 3.2 aufgrund der unterschiedlichen Koeffizientenempfindlichkeit im Durchlaß und Sperrbereich getroffen.

Approximation	$n_{\min}$	$C_{\min}$	$C_{\max}$
Potenz- verhalten	$\frac{\lg (\Delta_2 / \Delta_1)}{\lg v_S}$	$\Delta_2 v_S^{-n}$	Δ_1
Tschebyscheff- Durchlaßbereich	$\frac{\operatorname{ar} \cosh \Delta_2 / \Delta_1}{\operatorname{ar} \cosh v_S}$	$\frac{\Delta_2}{\cosh (n \operatorname{ar} \cosh v_S)}$	Δ_1
Tschebyscheff- Sperrbereich	$\frac{\operatorname{ar} \cosh \Delta_2 / \Delta_1}{\operatorname{ar} \cosh v_S}$	Δ_2	$\Delta_1 \cosh (n \cdot \operatorname{ar} \cosh v_S)$
Cauer	$\frac{K_O(\kappa^2) \cdot K_O'(\kappa_1^2)}{K_O'(\kappa^2) \cdot K_O(\kappa_1^2)}$	$\Delta_2 \Delta$	Δ_1 / Δ

Tabelle 2.1: Notwendiger Filtergrad und Grenzen der Entwurfskonstanten C bei unterschiedlicher Approximation mit den Beziehungen:

$$\kappa^2 = 1/v_S; \quad \kappa_1^2 = \Delta_1/\Delta_2 \quad ;$$

$$K_O(\kappa^2) = \text{vollständiges elliptisches Integral} \\ 1. \text{ Gattung}$$

$$K_O'(\kappa^2) = K_O(\sqrt{1-\kappa^4})$$

$$\Delta = \Delta(\kappa^2, n)$$

Für das in Bild 2.5a gezeigte Toleranzschema werden die notwendigen Filtergrade $n_{\min}$ und n in Tabelle 2.2 zum Vergleich zusammengestellt. Da die Cauer-Lösung optimal ist; ergibt sich für sie der kleinste Filtergrad.

Für die Bewertung des schaltungstechnischen Realisierungsaufwandes sind neben dem Filtergrad noch weitere Gesichtspunkte (z.B. Anzahl der notwendigen Multiplizierer) zu berücksichtigen. Ein diesbezüglicher Vergleich wird in Abschnitt 3.3 durchgeführt.

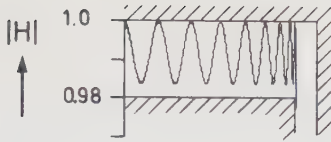
	$n_{\min}$	n
Potenz	57,30	58
Tscheby. Durchlaß.	16,47	17
Tscheby. Sperrber.	16,47	17
Cauer	7,99	8

Tabelle 2.2: Notwendige Filtergrade für Beispiele nach Bild 2.5a.

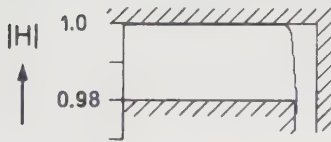
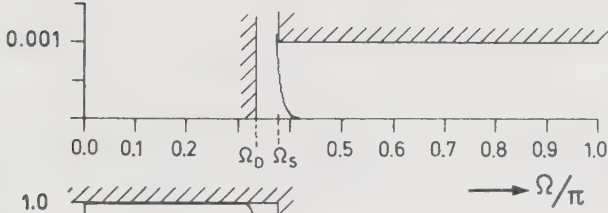
Die Berechnung der Pol- und Nullstellen der Übertragungsfunktionen bei unterschiedlicher Approximation kann nach [1], [2] erfolgen. Eine geschlossene Darstellung der Cauer-Lösung wird darüber hinaus in [51] gegeben.

Für das Toleranzschema nach Bild 2.5 sind die Beträge der Übertragungsfunktion bei Tschebyscheff- und Cauer-Approximation im Bild 2.6 eingetragen. Es wird jeweils der minimale Filtergrad gewählt. Das vorgelegte Toleranzschema kann gerade noch mit einem Cauer-Filter 8. Grades bei sehr kleinem Spielraum erfüllt werden. Für die beiden anderen Approximationsarten wird die Konstante C so gewählt, daß das Toleranzschema bei ungerundeten Koeffizienten gleichmäßig ausgenutzt wird.

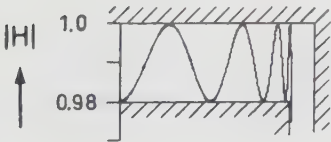
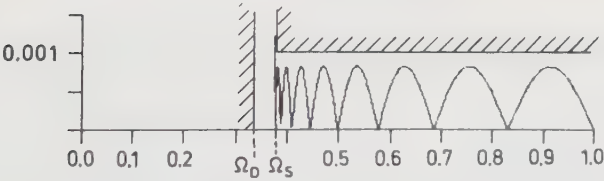
Die zugehörigen Datensätze sind dem Anhand zu entnehmen.



Tschebyscheff - V.
Durchlaßbereich



Tschebyscheff - V.
Sperrbereich



Cauer

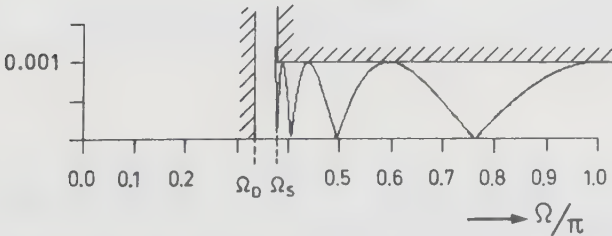


Bild 2.6: Beispiele zur unterschiedlichen Approximation des gleichen Toleranzschemas.

2.3 Unterschiedliche Darstellung und Quantisierung der Koeffizienten

Wie bereits einleitend bemerkt, ist es zur Realisierung eines digitalen Filters notwendig, die Koeffizienten auf eine endliche Wortlänge zu begrenzen [1], [6]. Die Wahl der Darstellung hängt dabei stark von der Art der Realisierung des Filters ab. Für den Entwurf eines festverdrahteten Filters wird man gegebenenfalls eine Darstellung wählen, die den Besonderheiten des jeweiligen Koeffizientensatzes gerecht wird. Soll jedoch das zu entwerfende Filter auf einem bestehenden einstellbaren Filtersystem (z.B. [38], [39]) betrieben werden, so wird man sich bei der Darstellung der Koeffizienten nach den vorgegebenen Möglichkeiten richten müssen.

2.3.1 Allgemeine Zahlendarstellungen der Koeffizienten

Geht man generell von einer seriellen Verarbeitung der zu filternden Daten aus [1], [38], [39], so eignet sich für einen günstigen Multiplikationsablauf die Koeffizienten-Darstellung nach Vorzeichen und Betrag. Erfolgt die Beschreibung des Betrags durch eine vereinfachte, d.h. "nicht normalisierte" Gleitkomma-Darstellung im dualen Zahlensystem, so kann man den verschiedensten Anforderungen an die Koeffizienten-Darstellung gerecht werden.

Die Aufteilung in der allgemeinen Form zeigt Bild 2.7. Unter der Voraussetzung, daß ein maximaler Zahlenbereich

$$|z| < 2^p$$

festgelegt ist, kann der Exponent auf negative Zahlen beschränkt bleiben. Die Angabe seines Vorzeichens ist somit für den Exponenten nicht erforderlich.

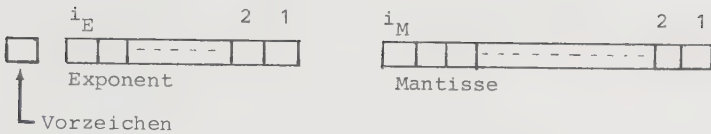


Bild 2.7: Allgemeine Darstellung der Koeffizienten

Mit einer Mantissenlänge i_M und einer Exponentenlänge i_E erhält man die Gesamtwortlänge des Koeffizienten

$$w_c = i_M + i_E + 1 \quad (2.20)$$

Der Betrag eines Koeffizienten x ist in dieser Darstellung

$$|x| = 2^{p+e-i_M-1} \cdot \sum_{i=1}^{i_M} \xi_{M_i} 2^i$$

mit dem Exponenten

$$e = - \sum_{i=1}^{i_E} \xi_{E_i} 2^{i-1} \quad (2.21)$$

Die Wahl des für die Darstellung maximalen Zahlenbereichs Z hängt von der für die Realisierung verwendeten Struktur und somit vom Zahlenbereich der Koeffizienten ab; er kann für verschiedene Koeffizienten verschieden sein.

Die kleinste Quantisierungsstufe der Koeffizienten berechnet sich je nach verwendetem Zahlenbereich zu

$$Q_c = 2^{p + e_{\min} - i_M} = e^{-q_c} \quad (2.22)$$

mit

$$e_{\min} = -(2^{i_E} - 1)$$

Durch den Verzicht auf eine Normalisierung der Mantisse entspricht die betragsmäßig kleinstmögliche Zahl ebenfalls der kleinsten Quantisierungsstufe Q_c

Durch die hier beschriebene eingeschränkte Gleitkomma-Darstellung, die im folgenden mit "Pseudo-Gleitkomma"-Darstellung bezeichnet ist, wird eine Einbeziehung verschiedener Realisierungsformen in ein allgemeines Entwurfsprogramm möglich. Wählt man speziell den Exponenten $e = 0$, so geht die Gleitkomma-Darstellung in die Festkomma-Darstellung über. Auch eine gemeinsame Verarbeitung des Exponenten für mehrere Koeffizienten wird möglich.

Eine weitere Variante der Koeffizientendarstellung für Blöcke zweiten Grades in der ersten oder zweiten kanonischen Form [1], wie sie im Abschnitt 3 zur Optimierung der Kaskadenschaltung verwendet werden, ist die Darstellung durch Differenzen. Da die Nennerkoeffizienten in (2.9) aus Stabilitätsgründen höchstens in den Bereichen

$$-2 < c_1 < 2$$

$$-1 < c_0 < 1$$

liegen können, kann man für Pole in der Umgebung von $z_\infty = 1$, also

$$c_1 \approx -2 \quad \text{und} \quad c_0 \approx 1,$$

auch durch die "Pseudo-Gleitkomma"-Darstellung keine Verbesserung der Koeffizientenfehler erreichen. Die Darstellung

$$c_1 = c_1' + 2 \operatorname{sgn} c_1 = c_1' - 2 \operatorname{sgn} c_1'$$

$$c_0 = 1 - c_0'$$

führt jedoch auf neue Koeffizienten c_1' , $c_0' \ll 1$. Sie lassen sich dann durch die "Pseudo-Gleitkomma"-Darstellung bei gleichem Aufwand wesentlich genauer vorgeben. Hierbei sind zwei zusätzliche Additionen nötig, die bei serieller Arbeitsweise keinen nennenswerten Mehraufwand bedeuten.

Bild 2.8 zeigt den Signalflußgraphen dieser modifizierten 1. kanonischen Form.

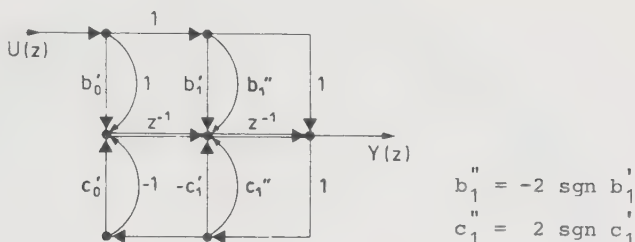
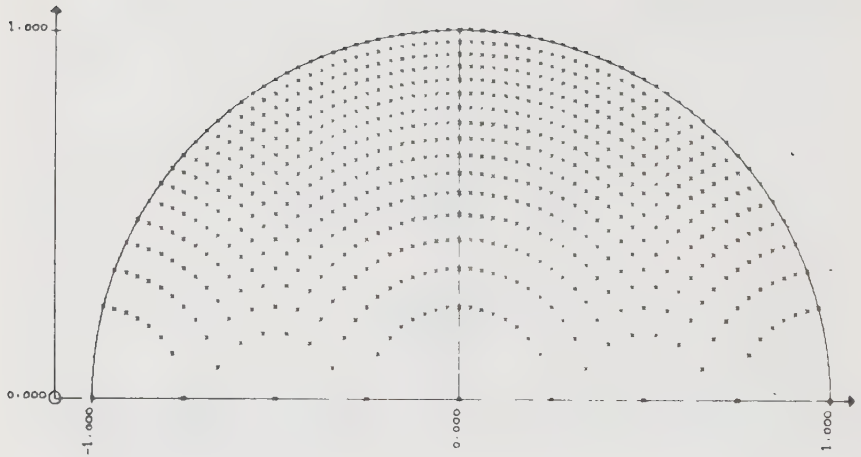


Bild 2.8: Modifizierte 1. kanonische Form.

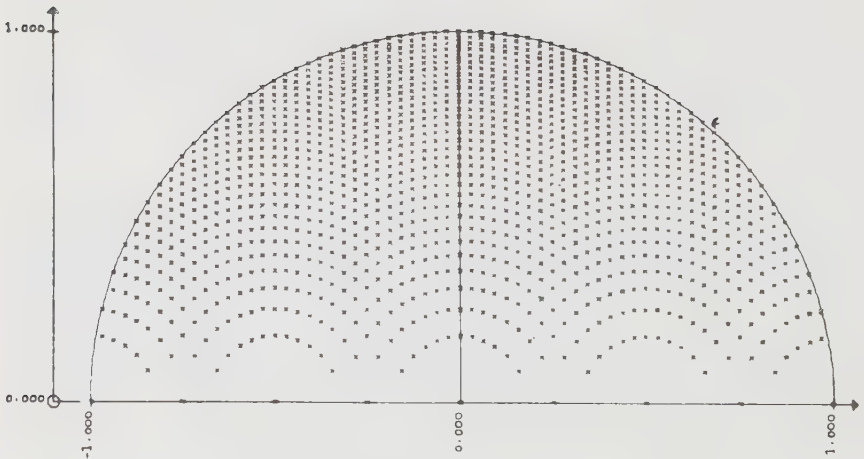
Die möglichen Pollagen verschiedener Koeffizientendarstellungen werden in den Bildern 2.9 und 2.10 für eine Gesamtwortlänge von jeweils $w_C = 6$ Bit gezeigt. Für eine Festkomma-Darstellung mit $p=1$, $i_E=0$, $i_M=5$ und somit $q_C=4$ für c_1 und c_0 , d.h. Vorgabe eines gleichgroßen möglichen Zahlenbereiches für beide Koeffizienten, wird das Gitter der möglichen Pollagen im Bild 2.9a gezeigt. Bei Beachtung der Stabilitätsbedingung wird der mögliche Wertebereich für c_0 nicht ausgenutzt. Eine Darstellung von c_0 mit $p=0$, $i_E=0$, $i_M=5$ und somit $q_C=5$ bei gleichem Bereich für c_1 zeigt Bild 2.9b.

Die mögliche Verdichtung der Pollagen in der Umgebung von $z=\pm 1$ zeigt die Gleitkomma-Darstellung im Zusammenhang mit Differenzendarstellung (Bild 2.10a). Dabei wurde für beide Koeffizienten $i_E=2$ und $i_M=3$ gewählt. Die Festlegung des Zahlenbereichs erfolgte mit $p=1$ für c_1 und $p=0$ für c_0 . Günstig wirkt sich diese Darstellung für schmale Bandfilter mit Bereichsgrenzen in der Gegend von $\Omega = \pi/2$ aus (Bild 2.10b), wenn nur c_0 durch c_0' ersetzt wird.

Unabhängig von den hier angestellten Überlegungen wurden ähnliche Darstellungen von Lacroix [40] im Zusammenhang mit einer generellen Gleitkomma-Realisierung und von Burrus et. al. [41] zur Realisierung von Festkomma-Schaltungen mit Polen und Nullstellen in der Umgebung von $z = 1$ angegeben.



a)



b)

Bild 2.9: Pollagen bei Festkomma-Darstellung

a) $i_E = 0$, $i_M = 5$, $p = 1$

b) $i_E = 0$, $i_M = 5$, $p = 1$ für c_1
 $p = 0$ für c_0

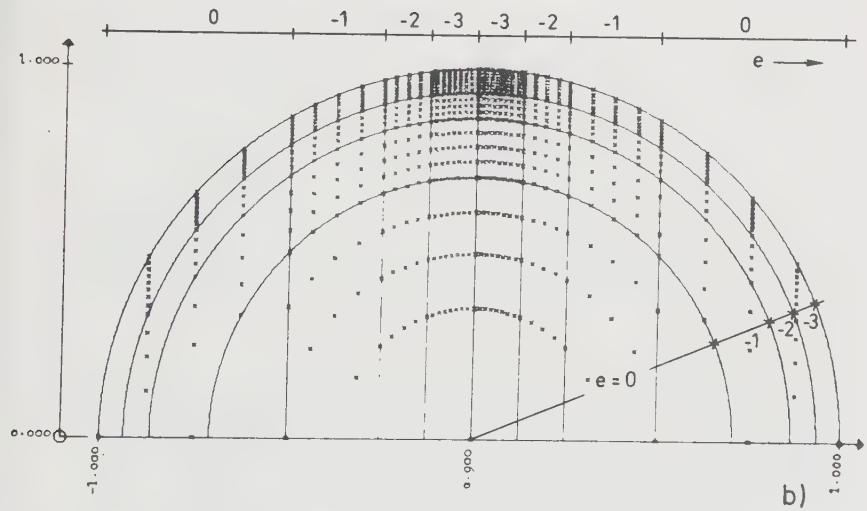
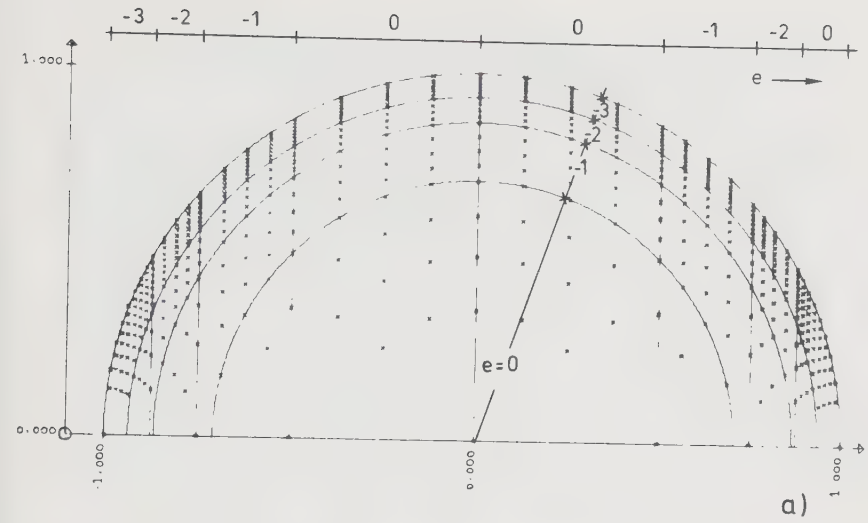


Bild 2.10: Pollagen bei Gleitkomma- und Differenzen-Darstellung, $i_E = 2$, $i_M = 3$

a) $c_0 = 1 - c'_0$ ($p=0$); $c_1 = c'_1 - 2 \operatorname{sgn} c'_1$ ($p=1$)
 c'_0 und c'_1 in Gleitkomma

b) $c_0 = 1 - c'_0$ ($p=0$); c_1 normal ($p=1$)
 c'_0 und c_1 in Gleitkomma

2.3.2 Bemerkungen zur Realisierung der verschiedenen Koeffizienten-Darstellungen

Die Wahl der Koeffizienten-Darstellung hängt stark von der technischen Realisierung des Filters ab. Betrachtet man z.B. eine festverdrahtete Schaltung, bei der zur Ausnutzung der maximalen, durch Filtergrad und Technologie festgelegten Abtastfrequenz sämtliche Multiplikationen getrennt realisiert werden (weder Block- noch Multiplizierer-Multiplex), so können die Koeffizienten fest verdrahtet werden. Die eigentlichen Multiplikationen beschränken sich auf die Mantisse, da der Exponent durch eine entsprechende zeitliche Verschiebung des Produktes realisiert werden kann. Ein Koeffizientenspeicher entfällt, man wird daher bei der Abschätzung des Realisierungsaufwandes in diesem Fall von der Mantissenwortlänge i_M ausgehen.

Werden bei verminderter Abtastfrequenz verschiedene Teile der Schaltung durch Multiplexbetrieb mehrfach ausgenützt, so müssen in jedem Fall Koeffizientenspeicher für Mantisse und Exponent vorgesehen werden. Eine flexible Exponenten-Darstellung benötigt darüber hinaus einen nicht zu vernachlässigenden schaltungstechnischen Aufwand zur Realisierung des Exponenten. In einer groben Näherung wird dieser Anteil am Gesamtaufwand proportional zur Exponentenwortlänge i_E gesetzt. Ebenso geht eine flexible Differenzen-Darstellung bei der Aufwandsabschätzung mit in die Bewertung ein.

Eine Vereinfachung ergibt sich, wenn eine einheitliche oder zumindest für verschiedene Koeffizienten gemeinsame Exponenten-Darstellung verwendet wird. So wurde z.B. in dem in [38], [39] beschriebenen rekursiven Filter in Kaskadenform mit flexiblen Koeffizienten eine einheitliche Darstellung für alle Zählerkoeffizienten und für die gleichartigen Nennerkoeffizienten gewählt. Die entsprechenden Exponenten bzw. Zahlenbereiche sind durch Schalter einstellbar.

Bei der "Pseudo-Gleitkomma"-Darstellung mit Exponenten $e < 0$ ist zu berücksichtigen, daß bei der Realisierung zusätzliche Taktzeiten für die Abarbeitung des Exponenten benötigt werden. Die Gesamtarbeitszeit richtet sich nach der minimalen Quantisierung der Koeffizienten. So benötigt zum Beispiel eine Festkomma-Darstellung mit $w_C = i_M + 1 = 16$ Bit die gleichen Zeiten wie eine "Pseudo-Gleitkomma"-Darstellung mit $i_M = 12$ Bit Mantisse und einen Exponenten $e \geq -(2^2 - 1) = -3$ ($i_E = 2$).

3. Entwurf selektiver digitaler Filter minimalen Koeffizientenaufwands bei Kaskadenschaltung

Dem Entwurf digitaler Filter minimalen Koeffizientenaufwands muß eine genaue Empfindlichkeitsanalyse der verschiedenen Strukturen vorausgehen. Untersuchungen wurden in grundlegenden Arbeiten von Knowles et. al. in [42], [43] von Rader und Gold in [44], Avenhaus und Schüßler in [22], [6] sowie von Fettweis in [45] durchgeführt. Aufbauend auf linearen Signalflußgraphen und deren Matrixbeschreibung wird in [46], [3] von Crochiere eine umfassende Darstellung einer allgemeinen digitalen Netzwerkanalyse einschließlich einer Untersuchung der Empfindlichkeit gegeben. Daneben werden allgemeine Analyseprogramme in [48] und [49] beschrieben. Eine zusammenfassende Darstellung des Standes der Untersuchungen mit weiteren Literaturhinweisen wird von Crochiere und Oppenheim in [47] gegeben.

Aufgrund einer Empfindlichkeitsanalyse fand Avenhaus, daß die Kaskadenschaltung von Blöcken 2. Grades in der ersten oder zweiten kanonischen Form für die Realisierung höhergradiger selektiver digitaler Filter besonders geeignet ist [6]. Insbesondere zeigt sich, daß die Kaskadenschaltung im Sperrbereich wesentlich weniger empfindlich ist als die entsprechende Parallelschaltung. Die geringe Empfindlichkeit im Sperrbereich läßt auf relativ unempfindliche Zählerkoeffizienten der Übertragungsfunktion schließen. Die unterschiedliche Empfindlichkeit der Übertragungsfunktion im Durchlaß- und Sperrbereich kann bei einer geeigneten Ausnutzung des Toleranzschemas bei ungerundeten Koeffizienten zur Minimierung der Koeffizientenwortlänge verwendet werden [1]. Eine nach Durchlaß- und Sperrbereich getrennte Untersuchung der Koeffizientenempfindlichkeit und die Bestimmung des maximalen Fehlers ist hierzu notwendig [1], [6], [22].

3.1 Untersuchung der Koeffizientenempfindlichkeit im Durchlaß- und Sperrbereich

3.1.1 Bestimmung des maximalen Fehlers in Abhängigkeit von der Koeffizientenwortlänge

Zur Untersuchung des im Durchlaß- und Sperrbereich unterschiedlichen Fehlerverhaltens werden im Gegensatz zu [6] für beide Bereiche getrennte Fehlerfunktionen eingeführt, die aber in Anlehnung an [6] definiert werden [37]. Der Index w weist auf die Koeffizienten endlicher Wortlänge hin.

Man erhält für den Durchlaßbereich

$$\epsilon_{Dw}(\Omega) = \frac{1}{\delta_D} \frac{\max |H_w| - |H_w(\Omega)|}{\max |H_w|} ; \text{ für } 0 \leq |\Omega| \leq \Omega_D , \quad (3.1a)$$

und im Sperrbereich

$$\epsilon_{Sw}(\Omega) = \frac{1}{\delta_S} \frac{|H_w(\Omega)|}{\max |H_w|} ; \quad \text{für } \Omega_S \leq |\Omega| \leq \pi . \quad (3.1b)$$

Dabei bedeuten δ_D , δ_S , Ω_D , Ω_S die Toleranzgrenzen entsprechend Bild 2.5 und $H_w(\Omega)$ die Übertragungsfunktion mit gerundeten Koeffizienten der Wortlänge w . ϵ_{Dw} , $\epsilon_{Sw} \leq 1$ ist zur Erfüllung der Forderungen nötig. Je nach Darstellung berechnet sich die Quantisierung der Koeffizienten nach Gl. 2.22. Für die hier folgenden Beispiele wird stets von einem Zahlenbereich $|z| < 2^p = 2$, also $p = 1$ und der Festkomma-Darstellung ($i_e = 0$) für die Koeffizienten ausgegangen. Es ergibt sich also

$$q_C = w_C^{-2}$$

Da der Aussteuerungsbereich stets auf Werte $H_w(\Omega) < 1$ für alle Ω beschränkt bleiben soll (siehe Abschnitt 4.3), wird eine Normierung des Fehlers auf

$$\max |H_W| = \max |H_W(\Omega)|; \quad 0 \leq |\Omega| \leq \pi \quad (3.2)$$

durchgeföhrt. Als Beispiel werden in Bild 3.1 die Fehlerfunktionen für einen Cauer-Tiefpaß 8. Grades berechnet, der das nach Bild 2.5 vorgesehene Toleranzschema gerade erfüllt. Die Abhängigkeit von der Frequenz Ω ist in Bild 3.1 für die Wortlängen $w_c = 17$ und $w_c = 11$ dargestellt. Offenbar sind $w_c = 11$ Bit im Durchlaßbereich nicht ausreichend, während sich im Sperrbereich die Reduzierung von 17 auf 11 Bit nicht auswirkt.

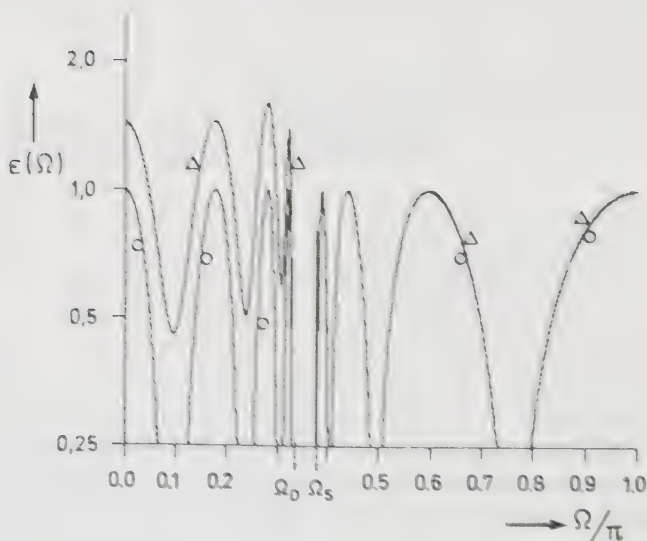


Bild 3.1: Fehlerfunktion beim Cauer-Tiefpaß 8. Grades in der Kaskadenschaltung

○ ○ ○ $w_c = 17$

△ △ △ $w_c = 11$

Verwendet man nun für die Approximation des gleichen Toleranzschemas ein Filter mit höherem Grad, z.B. $n = 10$, so sieht man in Bild 3.2, daß das Toleranzschema mit wesentlich geringe-

rer Wortlänge befriedigt werden kann. Auch hier wurde von einer gleichmäßigen Ausnutzung der Toleranzschemata im Durchlaß- und Sperrbereich bei ungerundeten Koeffizienten ausgegangen.

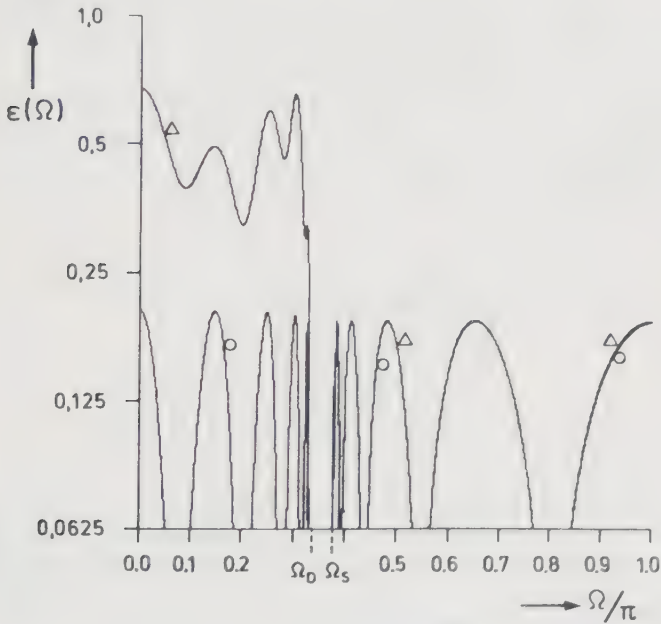


Bild 3.2: Fehlerfunktion beim Cauer-Tiefpaß 10. Grades
(gleichmäßige Ausnutzung bei $w_c = \infty$)
○ ○ $w_c = 17$, △ △ $w_c = 11$

Das Beispiel bestätigt die in [6] gemachte Aussage, nach der man im Sperrbereich einen wesentlich geringeren Fehler $\epsilon_{S|w=11} < 0,18$ erhält als im Durchlaßbereich ($\epsilon_{D|w=11} < 0,67$). Zur weiteren Beurteilung der gefundenen Approximation wird nun der im jeweiligen Toleranzbereich maximale Fehler verwendet

$$\max \epsilon_{Dw} = \max \{ \epsilon_{Dw}(\Omega) \}; \quad 0 \leq |\Omega| \leq \Omega_D, \quad (3.3a)$$

$$\max \epsilon_{Sw} = \max \{ \epsilon_{Sw}(\Omega) \}; \quad \Omega_S \leq |\Omega| \leq \pi. \quad (3.3b)$$

Die Abhängigkeit der maximalen Fehler $\max \epsilon_{Dw}$, $\max \epsilon_{Sw}$ von der verwendeten Wortlänge bzw. von der Quantisierung der Koeffizienten zeigt nun Bild 3.3 für die Tiefpässe 8. und 10. Grades.

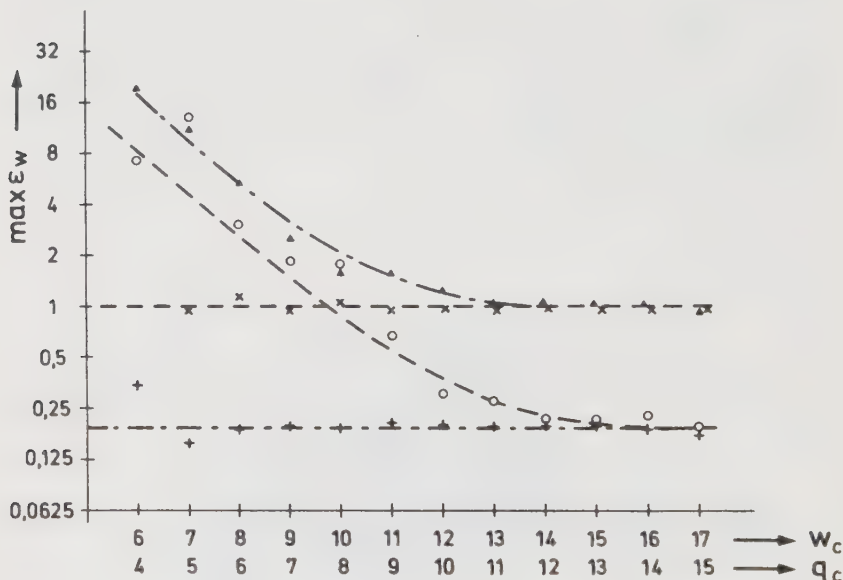


Bild 3.3: Abhängigkeit der Fehler $\max \epsilon_{Dw}$, $\max \epsilon_{Sw}$ von der Wortlänge w_c (Beispiele:

TP 8. Grades ▲ DB, x SB,

TP 10. Grades ○ DB, + SB)

Die zur Realisierung notwendige Wortlänge ist das Maximum der Wortlängen in den einzelnen Bereichen. Einen Ausgleich der notwendigen Wortlängen kann man durch eine unterschiedliche Grundausnutzung der Bereiche, d.h. durch eine geeignete Wahl der Entwurfskonstanten C nach Gl. (2.12) und Gl. (2.17) erzielen [1]. So wird z.B. bei einem C in der Nähe der unteren Grenzen der Durchlaßbereich bei unbegrenzter Wortlänge nur wenig ausgenutzt und der freie Toleranzbereich kann für den durch Rundung entstehenden zusätzlichen Fehler verwendet werden.

3.1.2 Abhängigkeit des maximalen Fehlers von der Entwurfskonstanten C

Durch eine systematische Untersuchung des maximalen Fehlers in Abhängigkeit von der Entwurfskonstanten C wird an einem Beispiel gezeigt, daß eine Optimierung der Konstanten im Hinblick auf eine möglichst geringe Gesamtwortlänge möglich ist [37]. Zunächst werden für die unbegrenzte Wortlänge $w_c = \infty$ (praktisch: Rechnergenauigkeit!) die maximalen Fehlerwerte entsprechend den Gln. (3.3) bestimmt. Es gilt $\max |H_\infty| = 1$. Mit $|H_\infty(\Omega)| = |H(\Omega)|$ nach Gl. (2.12) und den Werten $C_{\min}$ und $C_{\max}$ nach Tab. 2.1 erhält man für die verschiedenen Approximationen

$$\max \epsilon_{D_\infty} = \frac{1}{\delta_D} \left(1 - \sqrt{\frac{C_{\max}^2}{C_{\max}^2 + C^2 \Delta_1^2}} \right), \quad (3.4a)$$

$$\max \epsilon_{S_\infty} = \frac{1}{\delta_S} \frac{C_{\min}}{\sqrt{C_{\min}^2 + C^2 \Delta_2^2}} \quad (3.4b)$$

mit Δ_1, Δ_2 nach Gln. (2.13).

Speziell erhält man für die im Beispiel verwendete Cauer-Lösung

$$\max \epsilon_{D_\infty} = \frac{1}{\delta_D} \left(1 - \sqrt{\frac{1}{1 + C^2 \Delta^2}} \right),$$

$$\max \epsilon_{S_\infty} = \frac{1}{\delta_S} \frac{\Delta}{\sqrt{\Delta^2 + C^2}}.$$

Für das Beispiel eines Tiefpasses 10. Grades, das entsprechend Bild 3.2 das vorgegebene Toleranzschema bei ungerundeten Koeffizienten nur teilweise ausnutzt, werden nun die maximalen Fehler bei unterschiedlicher Wahl der Entwurfskonstanten C im Bild 3.4 gezeigt. Als Quantisierung wird $Q_C = 2^{-9}$ verwendet. Das entspricht bei einheitlicher Zahlendarstellung ($p = 1$) einer Koeffizientenwortlänge $w_c = 11$. Zum Vergleich werden im Bild 3.4 zusätzlich die maximalen Abweichungen $\max \epsilon_{D_\infty}$ und

$\max \epsilon_{S\infty}$ von der Wunschfunktion eingezeichnet. Die Bedeutung der Grenzkurven $\max \epsilon_w \Big|_{95\%}$ wird im nachfolgenden Abschnitt erklärt.

Die untere und obere Grenze der Konstanten C ist stets eine Funktion der charakteristischen Größen des Toleranzschemas und des gewählten Filtergrades. Um gegebenenfalls einen Vergleich mit anderen Entwurfsbeispielen durchführen zu können, erscheint die Einführung einer Normierung sinnvoll. Entsprechend den Gln. (3.5) erfolgt die Abbildung des Bereiches

$$C_{\min} \leq C \leq C_{\max}$$

auf einen Bereich

$$0 \leq c \leq 1.$$

Es ist

$$C = C_{\min} \cdot \left(\frac{C_{\max}}{C_{\min}} \right)^c \quad (3.5a)$$

oder

$$c = \frac{\lg C - \lg C_{\min}}{\lg C_{\max} - \lg C_{\min}}. \quad (3.5b)$$

Speziell für den Fall der Cauer-Approximation ergibt sich

$$C = \Delta_2 \Delta \left(\frac{\Delta_1}{\Delta_2} \cdot \frac{1}{\Delta^2} \right)^c$$

oder

$$c = \frac{\lg C - \lg \Delta_2 \Delta}{\lg \Delta_1 / \Delta - \lg \Delta_2 \Delta}$$

Wie man sieht, liegt der maximale Fehler bei begrenzter Wortlänge im Sperrbereich sehr nahe der Grenzkurve für unbegrenzte Koeffizienten. Hingegen sind die Abweichungen im Durchlaßbereich wesentlich größer und sehr unregelmäßig. Dennoch erkennt man eine gegenläufige Abhängigkeit der maximalen Fehler im Durchlaß- und Sperrbereich. Ein klares Minimum des Gesamtfehlers liefert diese experimentelle Untersuchung jedoch nicht. Durch

die Abschätzung einer oberen Schranke für die beiden Fehlerwerte können die gegenläufigen Abhängigkeiten jedoch weiter verdeutlicht werden.

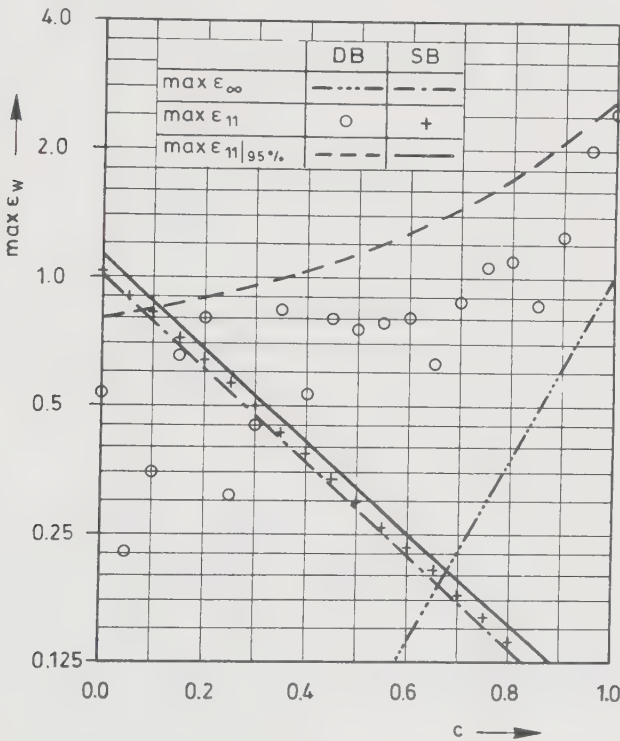


Bild 3.4: Fehler $\max \epsilon_w$ in Abhängigkeit von der Konstanten c ,
 $n = 10$, $\Omega_D = \pi/3$, $\Omega_S = 1.18$, $\delta_D = 0,02$, $\delta_S = 0,001$
 $w_C = 11$, $Q_C = 2^{-9}$.

Die Empfindlichkeit der Übertragungsfunktion im Durchlaßbereich wird besonders durch die Lage der Pole beeinflusst. Bild 3.5 zeigt die Polverschiebungen bei einer Veränderung der Entwurfskonstanten c und damit den Variationsbereich möglicher Lösungen. Man sieht, wie sich der Abstand vom Einheitskreis mit

zunehmendem c verringert. Gleichzeitig entfernen sich die Pole weiter von den Nullstellen, mit der Folge, daß der Maximalwert der Übertragungsfunktion im Sperrbereich geringer wird. Die Nullstellen selbst sind von c unabhängig.

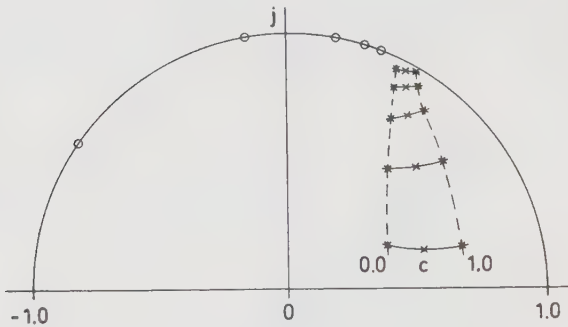


Bild 3.5: Veränderung der Pollagen in Abhängigkeit von c .

3.1.3 Abschätzung einer theoretischen Fehlergrenze

Die Fehlerwerte $\max \epsilon$ streuen besonders stark für den Durchlaßbereich. Durch die Abschätzung einer theoretischen Fehlergrenze wird gezeigt, daß es bei einer hinreichenden Koeffizientenwortlänge einen Bereich der Entwurfskonstanten c gibt, in dem sowohl im Durchlaß- wie im Sperrbereich das Toleranzschema mit großer Sicherheit eingehalten werden kann.

Von Avenhaus wird in [6] eine Abschätzung der erforderlichen Koeffizientenwortlänge anhand einer Fehlerstatistik durchgeführt. Dabei werden die gerundeten Koeffizienten a_{wv} mit

$v = 1(1) n_a$ einer Übertragungsfunktion $H_w(\Omega)$ als Zufallsvariable eines Ensembles von Übertragungsfunktionen $\tilde{H}(\Omega)$ gewertet. Die zugehörigen Koeffizienten $\tilde{a}_v = a_v + \Delta a_v$ liegen in einem Intervall $\pm \frac{Q}{L}$ um den ungerundeten Koeffizienten a_v . Somit ist auch die Übertragungsfunktion mit gerundeten Koeffizienten $H_w(\Omega)$ in dem Ensemble $\tilde{H}(\Omega)$ enthalten. Nach [6] kann das Ensemble der zugehörigen Leistungsübertragungsfunktionen $\tilde{P}(\Omega)$ mit Hilfe der Empfindlichkeiten der zur Übertragungsfunktion $H(\Omega)$ gehörenden Leistungsübertragungsfunktion $P(\Omega)$ berechnet werden. Es gilt

$$P(\Omega) = H(\Omega) H^*(\Omega) = |H(\Omega)|^2 \quad (3.6)$$

$$\tilde{P}(\Omega) = P(\Omega) + \sum_{v=1}^n a_v \frac{\partial P(\Omega)}{\partial a_v} \Delta a_v. \quad (3.7)$$

Nimmt man weiter an, daß die einzelnen Koeffizientenfehler gleichverteilt und voneinander unabhängig sind, so ergibt sich der Mittelwert μ_P und die Streuung σ_P^2 der verschiedenen Leistungsübertragungsfunktion $\tilde{P}(\Omega)$ zu

$$\mu_P \approx \overline{\tilde{P}(\Omega)} \quad (3.8a)$$

$$\sigma_P^2 \approx \frac{Q^2}{12} \sum_{v=1}^n a_v^2 \left(\frac{\partial P(\Omega)}{\partial a_v} \right)^2 \quad (3.8b)$$

Entsprechend den in Bild 3.6 eingezeichneten Fehlerwerten kann nun folgende Abschätzung durchgeführt werden.

Für die Ausgangslösung $H(\Omega)$ gilt mit der gewünschten Übertragungsfunktion $H_w(\Omega)$

$$(|H_w(\Omega)| - \delta(\Omega) \cdot \max_{\epsilon_{\infty}})^2 \leq P(\Omega) \quad (3.9)$$

und entsprechend gilt für eine Fehlerschranke $\max \epsilon_w|_S$ (siehe Bild 3.6) bei gerundeten Koeffizienten, die mit einer Sicherheit S eingehalten wird

$$\begin{aligned} (|H_w(\Omega)| - \delta(\Omega) \cdot \max \epsilon_w|_S)^2 &\leq \tilde{P}(\Omega) = \\ &= P(\Omega) + \Delta \tilde{P}(\Omega). \end{aligned} \quad (3.10)$$

Für $\Delta \tilde{P}(\Omega)$ ergibt sich somit die Ungleichung

$$|\Delta \tilde{P}(\Omega)|_S = |\Delta \tilde{P}(\Omega)| \leq |(|H_W(\Omega)| - \delta(\Omega) \max \epsilon_\infty)^2 - (|H_W(\Omega)| - \delta(\Omega) \max \epsilon_W|_S)^2| \quad (3.11)$$

Setzt man die entsprechenden Größen für Durchlaß- und Sperrbereich in Gl. (3.11) ein, so erhält man

$$|\Delta \tilde{P}(\Omega)| \leq |(1 - \delta_D \max \epsilon_{D\infty})^2 - (1 - \delta_D \max \epsilon_{Dw}|_S)^2| \quad \text{für } 0 \leq |\Omega| \leq \Omega_D \quad (3.12a)$$

und

$$|\Delta \tilde{P}(\Omega)| \leq |\delta_S^2 (\max^2 \epsilon_{Sw}|_S - \max^2 \epsilon_{S\infty})| \quad \text{für } \Omega_S \leq |\Omega| \leq \pi \quad (3.12b)$$

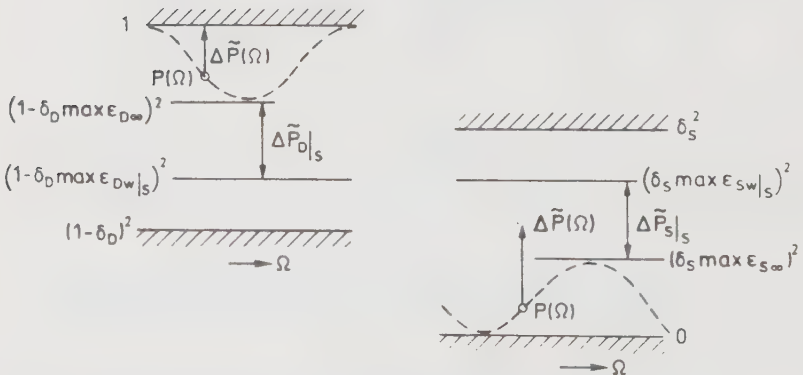


Bild 3.6: Toleranzgrenzen zur Abschätzung einer theoretischen Fehlergrenze

Unter der Voraussetzung $\mu_p = \tilde{P}(\Omega) \approx P(\Omega)$ kann aufgrund des Grenzwertsatzes von Lindenber-Feller [50] angenommen werden, daß sich die Verteilung $p(\tilde{P}(\Omega))$ mit steigender Variablenzahl einer Normalverteilung annähert. Eine Grenze für $|\Delta \tilde{P}(\Omega)|$ kann

somit abgeschätzt werden und man erhält für eine feste Frequenz Ω für mindestens 95 % der Übertragungsfunktionen des Ensembles

$$|\Delta\tilde{P}(\Omega)|_{95\%} \leq 2 \sigma_P(\Omega) \quad (3.13)$$

Bleibt die Normierung der Übertragungsfunktion $\tilde{H}(\Omega)$ auf den Maximalwert $\max |\tilde{H}(\Omega)|$ unberücksichtigt, so kann sich eine symmetrische Fehlerverteilung ausbilden. Die für die Abschätzung nach Gl. (3.13) getroffene Voraussetzung für den Mittelwert μ_P der Leistungsübertragungsfunktion ist erfüllt. Durch die Normierung der Übertragungsfunktion und der Fehlerfunktionen nach Gl. (3.1) entsteht bei der Berechnung der Leistungsübertragungsfunktionen $\tilde{P}(\Omega)$ eine "einseitig gestutzte" und somit unsymmetrische Verteilung der Leistungsfehler $\Delta\tilde{P}(\Omega)$. Je nach Lage der Ausgangslösung $P(\Omega)$ ist die Verteilung im Durchlaßbereich für positive und im Sperrbereich für negative Werte begrenzt.

Im allgemeinen führt die veränderte Verteilung zu veränderten Sicherheitsgrenzen. Zur Überprüfung der Sicherheitsgrenzen wurde daher die Verteilung der Fehlerwerte $\Delta\tilde{P}(\Omega)$ bei Variation der Koeffizienten für verschiedene Werte Ω an einigen Beispielen berechnet. Die experimentellen Ergebnisse zeigten, daß die 95 %-Grenze nach Gl. (3.13) auch bei der schiefen Verteilung eingehalten wird.

Zur Bestimmung der maximalen Fehlerschranke wird nun der maximale Wert der Streuung in dem jeweiligen Frequenzbereich verwendet. Mit Gl. (3.6) und (3.8b) erhält man

$$\max \sigma_P = \frac{\Omega_C}{2\sqrt{3}} \max_{\Omega} \sqrt{\sum_{v=1}^n a_v \left(\frac{\partial P(\Omega)}{\partial a_v} \right)^2} \quad , \quad (3.14)$$

$$\forall \Omega \in [0, \pi] \quad .$$

Es sei:

$$\max \sigma_{P_D} = \max \sigma_P \quad \text{für} \quad 0 \leq |\Omega| \leq \Omega_D$$

$$\max \sigma_{P_S} = \max \sigma_P \quad \text{für} \quad \Omega_S \leq |\Omega| \leq \pi$$

Durch Umformen der Gln. (3.12) erhält man mit Gl. (3.13)

$$\max \epsilon_{Dw} \Big|_{95\%} = \frac{1}{\delta_D} \left[1 - \sqrt{(1 - \delta_D \max \epsilon_{D\infty})^2 - 2 \max \sigma_{P_D}} \right] \quad (3.15a)$$

$$\max \epsilon_{Sw} \Big|_{95\%} = \frac{1}{\delta_S} \sqrt{\delta_S^2 \max \epsilon_{S\infty}^2 + 2 \max \sigma_{P_S}} \quad (3.15b)$$

Die Abhängigkeit der maximalen Streuungen von der Entwurfskonstante zeigt Bild 3.7. im logarithmischen Maßstab. Die gegenläufige Änderung im Durchlaß- und Sperrbereich unterstützt dabei die Wirkung der unterschiedlichen Grundausnutzung.

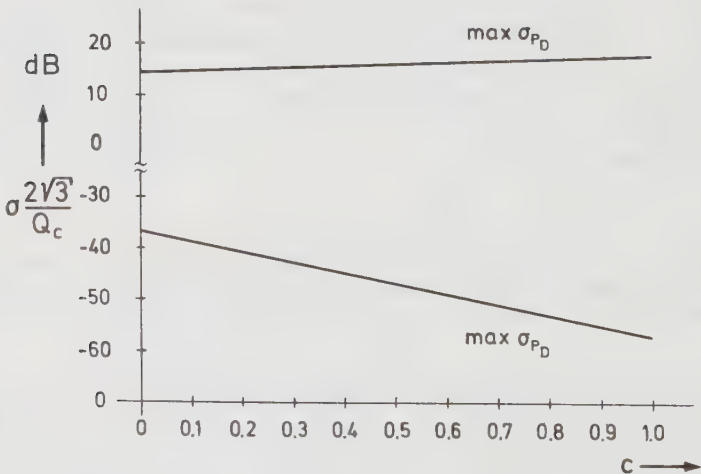


Bild 3.7: Maximale Streuung im Durchlaß- und Sperrbereich in Abhängigkeit von c am Beispiel eines Tiefpasses 10. Grades.

Zum Vergleich mit den tatsächlichen Fehlerwerten wurden die Grenzkurven für $w_c = 11$, entsprechend $Q_c = 2^{-9}$ berechnet und in das Bild 3.4 eingetragen. Man kann aus der Abschätzung schließen, daß die Wortlänge $w_c = 11$ mit großer Wahrscheinlichkeit ausreichend sein wird, wenn die Konstante c in einem Bereich $0.05 \leq c \leq 0.36$ gewählt wird.

Die Abhängigkeit der Grenzkurven von der verwendeten Wortlänge zeigt Bild 3.8. Man sieht, wie sich der Variationsbereich der Konstanten c für größere Wortlängen erweitert. Umgekehrt bedeutet dies, daß die beim Entwurf angestrebte minimale Koeffizientenwortlänge nur in einem stark eingeschränkten Bereich mit Sicherheit zum Ziel führt. Es erscheint daher notwendig, mittels eines Optimierungsalgorithmus eine geeignete Entwurfskonstante c zu bestimmen. In der Regel wird dabei die ermittelte notwendige Wortlänge geringer als der Schätzwert sein.

Bei den vorangehenden Untersuchungen der Kaskadenanordnung wurde stets davon ausgegangen, daß die zur Reduzierung des Rundungsrauschens (Abschnitt 5) notwendige Skalierung durch getrennte Vorfaktoren realisiert wird, die aufgrund der ohnehin durchgeführten Normierung (Gl. 3.2) keinen Einfluß auf die Empfindlichkeit haben. Bezieht man jedoch die Skalierung mit in die Zählerkoeffizienten ein, so beeinflussen diese ebenfalls die Empfindlichkeit. Besonders bei Skalierungsfaktoren $b_{2\lambda} < 1$ steigen die Empfindlichkeiten der Zählerkoeffizienten und somit gegebenenfalls der maximale Fehler im Sperrbereich. Eine Reihe von Optimierungsbeispielen hat diesen Sachverhalt bestätigt.

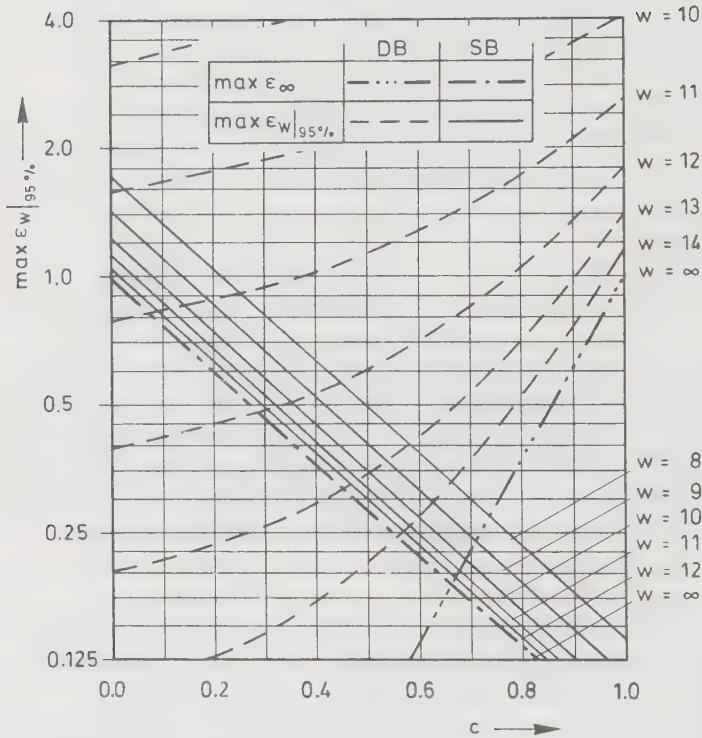


Bild 3.8: $\max \epsilon_w|_{95\%}$ in Abhängigkeit der Konstanten c

für das Beispiel des Tiefpasses 10. Grades,
Parameter: Wortlänge $w = w_c$; $Q_c = 2^{-(w_c-2)}$.

3.2 Optimierung der notwendigen Koeffizientenwortlänge durch Variation des Entwurfsparameters c

Die im vorausgehenden Abschnitt durchgeführten Untersuchungen des maximalen Fehlers der Übertragungsfunktion bei gerundeten Koeffizienten lassen einen günstigen Bereich für die Wahl der Entwurfskonstanten c erkennen. Eine exakte analytische Eingrenzung ist jedoch wegen der hochgradig nichtlinearen Ab-

hängigkeit der maximalen Fehlerwerte (siehe Bild 3.4) nicht möglich. Mit Hilfe eines adaptiven Suchoptimierungsverfahrens wird es möglich einen günstigen Wert für c bei minimaler Koeffizientenwortlänge zu bestimmen. Ein dafür geeignetes Verfahren wird im nächsten Abschnitt beschrieben.

3.2.1 Adaptive Suchoptimierung

Die Variation der normierten Entwurfskonstanten c ist grundsätzlich in dem Bereich $0 \leq c \leq 1$ möglich. Die im Abschnitt 3.1 durchgeführten Untersuchungen lassen in vielen Fällen jedoch einen eingeschränkten Suchbereich, $c_{\min} \leq c \leq c_{\max}$, sinnvoll erscheinen. Nach Vorgabe eines geeigneten Startwerts c_1 wird die Optimierung wie folgt durchgeführt.

Allgemein werden im i -ten Schritt für $c = c_i$ die Koeffizienten des Filters berechnet und entsprechend der gewählten Koeffizientendarstellung umgeformt. Die maximalen Fehler $\max \epsilon_{Dw}$ im Durchlaßbereich und $\max \epsilon_{Sw}$ im Sperrbereich werden nun für verschiedene Koeffizientenwortlängen w_c bestimmt. Die Verkürzung von w_c wird dabei solange fortgesetzt, bis jeweils mindestens zwei aufeinander folgende Wortlängen geprüft sind, die das Toleranzschema erfüllen und zwei, die es nicht mehr erfüllen. Die zusätzliche Bestimmung der Randwerte wird notwendig, da bei einer starken Streuung der maximalen Fehlerwerte (siehe Bild 3.3) häufig kein scharfer Übergang zwischen den Wortlängen besteht, die das Toleranzschema erfüllen und denen, die es nicht erfüllen.

Die Wortlänge, die im i -ten Schritt das Toleranzschema gerade noch erfüllt, wird nun mit w_{c_i} bezeichnet. Zur Bestimmung der Entwurfskonstanten c_{i+1} des nachfolgenden Schritts wird nun die Entscheidungsfunktion $\Delta \epsilon_i$

$$\Delta \epsilon_i = \max_{\Delta c} \epsilon_{Dw} - \max_{\Delta c} \epsilon_{Sw} \quad (3.16)$$

eingeführt. Für die Wahl der entscheidenden Wortlänge $w_{\Delta\epsilon}$ sind zwei Fälle zu unterscheiden:

1. Wird bei der Wortlänge $w_c = w_{c_i} - 1$, die das Toleranzschema gerade nicht mehr erfüllt, nur von einem Fehlerwert die Schranke $\max \epsilon_w = 1$ überschritten, so besteht die Chance, daß im nachfolgenden Schritt ein weiterer Ausgleich der Fehlerwerte stattfinden kann und das Toleranzschema mit dieser Wortlänge im Durchlaß- und Sperrbereich eingehalten werden kann. Es wird also

$$w_{\Delta\epsilon_i} = w_{c_i} - 1$$

für die Bildung der Entscheidungsfunktion verwendet.

2. Werden jedoch für $w_c = w_{c_i} - 1$ beide Fehlerwerte gleichzeitig größer als eins, so besteht wenig Hoffnung auf eine weitere Reduzierung der Wortlänge, und der nachfolgende Schritt dient in erster Linie zum Ausgleich der unterschiedlichen Ausnutzung in den verschiedenen Toleranzbereichen. Für die Bestimmung der Entscheidungsfunktion wird daher

$$w_{\Delta\epsilon_i} = w_{c_i}$$

gewählt.

Aufgrund der starken Streuung der Fehlerwerte erscheint eine direkte Verwendung der Größe $\Delta\epsilon$ für die Bestimmung eines neuen Wertes c_{i+1} ungeeignet. Entsprechende Versuche haben keine günstigen Ergebnisse erbracht. Vielmehr wird entsprechend Gl. (3.16) nur die Richtung, in der c geändert werden soll, bestimmt.

Es gilt:

$$c_{i+1} < c_i \quad \text{für} \quad \Delta\epsilon_i > 0,$$

$$c_{i+1} > c_i \quad \text{für} \quad \Delta\epsilon_i < 0.$$

Der Wert von c_{i+1} selbst wird dann durch eine lineare Mittelung zwischen dem vorliegenden Wert c_i und den jeweiligen Randwerten des verwendeten Suchbereichs errechnet. Es sind dabei wiederum zwei unterschiedliche Suchbereiche zu unterscheiden: der ursprüngliche Bereich ($c_{\min}$, $c_{\max}$) und ein während der Optimierung selbständig eingegengter Bereich (c_N , c_P). Die Auswahl des Suchbereichs hängt von den vorausgegangenen Suchschritten ab und wird entsprechend dem Diagramm in Bild 3.9a gesteuert.

Bei der Untersuchung einer großen Anzahl von Optimierungsbeispielen hat sich der dort gezeigte heuristische Algorithmus als günstig erwiesen. An einem Beispiel zur Optimierung eines Tiefpasses mit Tschebyscheff-Verhalten im Sperrbereich (Toleranzschema nach Bild 2.5a) wird in Bild 3.9b der Optimierungsablauf, d.h. die Wahl der c_i und die Festlegung der Randwerte c_P und c_N verdeutlicht. Ausgehend von dem Startwert $c_1 = 0.75$ wird der Entwurf durchgeführt und die Entscheidungsfunktion gebildet. Man erhält $\Delta\epsilon_1 > 0$ und $c_2 = (c_1 + c_N)/2 = (c_1 + c_{\min})/2$. Gleichzeitig wird der Randwert c_P für $\Delta\epsilon_1 > 0$ gesetzt. Entsprechend wird beim 3. Schritt mit $\Delta\epsilon_3 < 0$ der Wert c_N festgelegt. Eine Korrektur der Werte c_P und c_N erfolgt dann, wenn sich bei den nachfolgenden Suchschritten bei minimaler Wortlänge w_{c_i} ein betragsmäßig kleinerer Fehlerwert $\Delta\epsilon_i$ ergibt.

Die Optimierung wird nun zunächst in dem sich selbständig verengenden Suchgebiet durchgeführt (Schritte 4-11). Um ein vorzeitiges Ende der Optimierung in einem lokalen Minimum zu vermeiden, erscheint es sinnvoll, entsprechend der Ablaufsteuerung in Bild 3.9a nach einer Häufung von Fehlerwerten mit gleichem Vorzeichen den eingegengten Bereich zu verlassen und eine erneute Suche im ursprünglichen Bereich vorzunehmen. So wird z.B. nach den Schritten 7-11 im Schritt 12 der Rand c_N überschritten. Ebenso erhält man nach den Schritten 12, 13, 14 eine globale Suche im Bereich $c > c_P$ mit dem Erfolg, daß die notwendige Gesamtwortlänge weiter verkürzt werden kann. Da das neue Minimum außerhalb des Bereiches (c_P , c_N) liegt, werden die Werte c_P und c_N in den nachfolgenden Schritten neu gesetzt.

Start: $S_1 = S_2 = F$

$F \hat{=}$ FALSE

$T \hat{=}$ TRUE

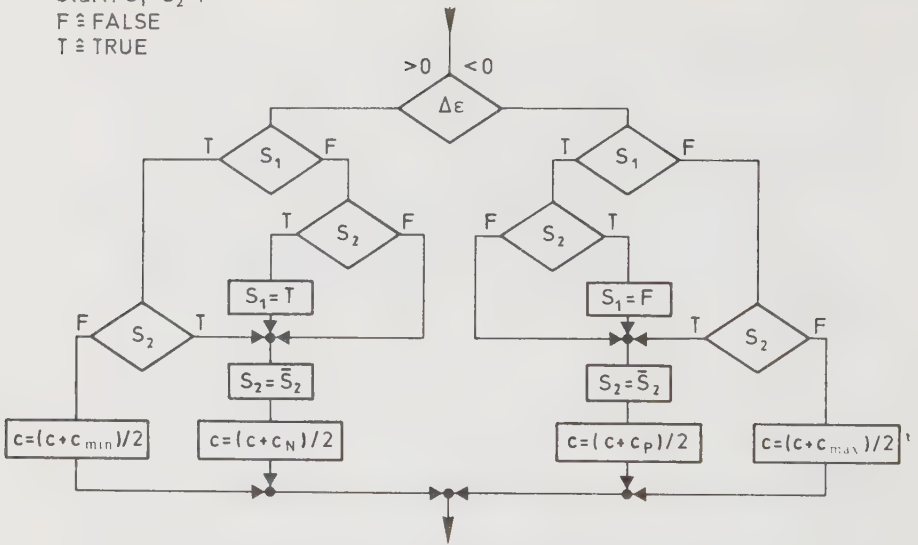


Bild 3.9a: Änderung der Entwurfskonstanten

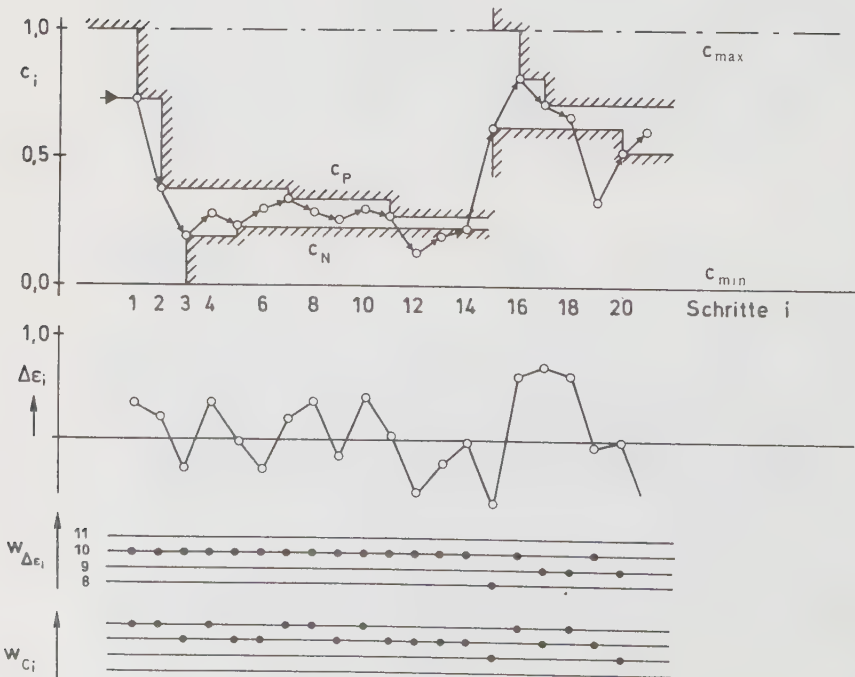


Bild 3.9b: Beispiel für einen Optimierungsablauf

In dem vorliegenden Beispiel wurde die Optimierung nach der vorgegebenen maximalen Anzahl an Suchschritten abgeschlossen. Als weiteres Abbruchkriterium kann eine maximale Anzahl aufeinanderfolgender erfolgloser Suchschritte festgesetzt werden.

3.2.2 Programmsystem zum optimierten Filterentwurf

Im Rahmen der vorliegenden Arbeit wurde ein umfassendes Entwurfs- und Optimierungsprogramm zum Entwurf selektiver rekursiver digitaler Filter erstellt. Ein vorgegebenes Toleranzschema kann dabei mit den verschiedenen in Abschnitt 2.2.2 beschriebenen Verfahren approximiert werden. Die mit Rechnergenauigkeit erstellten Koeffizienten der Übertragungsfunktion dienen als Ausgangslösungen für die Übertragungsfunktionen mit gerundeten Koeffizienten. Eine Optimierung der Koeffizienten für Kaskadenfilter mit Teilblöcken 2. Grades in der ersten und zweiten kanonischen Form (Bild 2.2) ist mit dem im vorhergehenden Abschnitt beschriebenen Algorithmus möglich. In Zusammenarbeit mit der Optimierung der Reihenfolge und Zuordnung (Abschnitt 5.3) kann eine Skalierung durch Einbeziehung der Vorfaktoren in die Zählerkoeffizienten erfolgen.

Eine ausführliche Beschreibung des Programmes DOREDI ($\hat{=}$ Design and Optimization of Recursive Digital Filters) wird in [51] gegeben.

3.2.3 Ausblick auf weitere Entwurfsmöglichkeiten

Eine Erweiterung des Programmsystems durch eine zusätzliche Umrechnung des Koeffizienten in die anderen Grundstrukturen, wie sie in Abschnitt 5.1 vorgestellt werden, ist möglich. In entsprechender Weise könnte auch die Optimierung der Parallelform nach Bild 2.3 mit in ein allgemeines Entwurfsprogramm einbezogen werden. Entsprechend den Untersuchungen in [6] zeigt jedoch die Parallelform keine so starken Unterschiede der maximalen Fehlerwerte im Durchlaß- und Sperrbereich bei gleichmäßiger Ausnutzung des Toleranzschemas mit ungerundeten Koeffizienten. Eine Optimierung des Entwurfsparameters c wird vermutlich bei dieser Struktur keine entsprechend guten Erfolge bringen. Genau-

ere Untersuchungen wurden jedoch im Rahmen der vorliegenden Arbeit nicht vorgenommen.

3.3 Minimierung des Koeffizientenaufwandes

Bereits in den Arbeiten [1], [6], [26] wurde gezeigt, daß die für die Realisierung notwendige Koeffizientenwortlänge in einigen Fällen durch Erweiterung des Filtergrades reduziert werden kann. Bei der Realisierung eines Filters auf dem am Institut für Nachrichtentechnik vorhandenen programmierbaren Gerät [39] kann in vielen Fällen durch Erhöhung des Filtergrades ein Toleranzschema eingehalten werden, das mit dem minimalen Grad wegen des Koeffizientenfehlers bereits verletzt wird. Bei vorhandenem Gerät, das im vorliegenden Fall $n = 32$ zuläßt, bedeutet die Filtergraderweiterung keinen zusätzlichen Aufwand. Hingegen wird beim Entwurf eines fest programmierten (verdrahteten) Filters für eine spezielle Anwendung der durch den zusätzlichen Filtergrad notwendige schaltungstechnische Mehraufwand berücksichtigt werden müssen. Daneben ist bei einem Multiplexbetrieb [1], [38] der Einfluß der Graderhöhung auf die maximale verarbeitbare Abtastfrequenz zu berücksichtigen.

3.3.1 Schaltungstechnische Bewertung der Koeffizientenwortlänge

Der schaltungstechnische Aufwand und damit die Realisierungskosten einer digitalen Schaltung hängen von einer großen Zahl z.T. nicht technischer Parameter ab und sind daher in einer allgemein gültigen Form kaum angebar. Abgesehen von den sich ändernden Bauteilekosten ermöglichen neue höherintegrierte Baugruppen oder eingeführte Baunormen gegebenenfalls eine kostenmäßig günstigere Auslegung der Schaltung. Darüber hinaus haben die geforderte Verarbeitungszeit und die sich daraus ergebende Entscheidung über die anzuwendende Technologie bzw. die Möglichkeit, Teile der Schaltung im Multiplexbetrieb aus-

zunutzen, einen wesentlichen Einfluß.

Für das Beispiel eines Kaskadenfilters mit Grundblöcken in der ersten kanonischen Form [38] wurde in [52] eine allgemeine Aufwandsformel, basierend auf den Bauteilekosten für verschiedene Filtertypen angegeben. In die Berechnungen gehen dabei die Parameter: notwendige Koeffizientenwortlänge w_c , notwendige Zustandsvariablenwortlänge w_s sowie Filtergrad n und möglicher Multiplexgrad ein. Entsprechende Aufwandsabschätzungen wurden in [53] für nichtrekursive Filter unterschiedlicher Struktur durchgeführt.

Im Rahmen der vorliegenden Arbeit erscheint eine absolute Bewertung des schaltungstechnischen Aufwands unmöglich. Vielmehr soll in sehr stark vereinfachter Form eine relative Bewertung der verschiedensten Approximationen eines vorgegebenen Toleranzschemas und deren unterschiedliche Realisierungen angegeben werden.

Eine Aufteilung des Gesamtaufwandsmaßes A in Teilfaktoren erleichtert dabei den Vergleich verschiedener Filterentwürfe für ein bestimmtes Toleranzschema.

$$A = A_C A_N A_S A_R \quad (3.17)$$

Die Bedeutung der einzelnen Teilaufwandsmaße wird im folgenden erläutert.

Geht man beim Vergleich verschiedener digitaler Filter davon aus, daß bei gegebener Technologie stets die maximale Abtast-rate erreicht werden soll, so müssen alle Multiplizierer getrennt realisiert werden. Unter der weiteren Annahme, daß der Aufwand für eine Multiplikation proportional zur Koeffizientenwortlänge w_c sein soll, erhält man bei der Kaskadenschaltung von Teilblöcken 2. Grades für einen geraden Filtergrad n und m_R Multiplikationen je Block ein Aufwandsmaß

$$A \sim \frac{n}{2} m_R w_c \quad (3.18)$$

Zum Vergleich verschiedener Filterentwürfe mit unterschiedlichen Realisierungen und gegebenenfalls auch unterschiedlichen Toleranzschemata wird der jeweils notwendige Aufwand auf einen theoretischen Mindestaufwand $A_{\min}$ bezogen.

$$A_{\min} \sim \frac{n_{\min}}{2} m_{R_{\min}} w_{C_{\min}} \quad (3.19)$$

Mit $n_{\min}$ wurde in Gl. (2.15) die theoretische Filtergradgrenze bezeichnet. Für $n \rightarrow n_{\min}$ wird das vorgeschriebene Toleranzschema bei unbegrenzter Wortlänge $w_C \rightarrow \infty$ gerade ausgenutzt. Bei hinreichender Filtergraderweiterung Δn

$$\Delta n = n - n_{\min} \quad (3.20)$$

nähert man sich andererseits einer unteren Grenze $w_{C_{\min}}$, die mit großer Sicherheit nicht mehr unterschritten werden kann. Beim Vergleich unterschiedlicher Realisierungen mit verschiedener Anzahl vollständiger Multiplizierer (siehe Abschnitt 5.3.4) wird mit $m_{R_{\min}}$ die minimal mögliche Anzahl bezeichnet.

Somit ergibt sich das Aufwandsmaß A_N zur Bewertung des Mehraufwands bei Filtergraderweiterung

$$A_N = \frac{n}{n_{\min}} = 1 + \frac{\Delta n}{n_{\min}} \quad (3.21)$$

und der relative Koeffizientenaufwand A_C bei einer einheitlichen Approximation und einer Realisierung mit der gleichen Grundstruktur

$$A_C = \frac{w_C}{w_{C_{\min}}} \quad (3.22)$$

Mit

$$A_{NC} = A_N A_C = \left(1 + \frac{\Delta n}{n_{\min}}\right) \frac{w_C}{w_{C_{\min}}} \quad (3.23)$$

erhält man somit ein Aufwandsmaß für die Koeffizientenoptimierung bei gleichzeitiger Filtergraderweiterung.

Zur Bewertung unterschiedlicher Strukturen wird jetzt noch ein

Bewertungsmaß A_R eingeführt. Es wird wieder auf eine rein theoretische, minimale Größe bezogen. Man erhält:

$$A_{R\kappa} = \frac{w_{C_{\min\kappa}} \cdot m_{R\kappa} \cdot n_{\min\kappa}}{w_{C_{\min}} \cdot m_{R_{\min}} \cdot n_{\min}} \quad (3.24)$$

mit den Werten

$$w_{C_{\min}} = \min_{\kappa} \{w_{C_{\min\kappa}}\}$$

$$m_{R_{\min}} = \min_{\kappa} \{m_{R\kappa}\}$$

$$n_{\min} = \min_{\kappa} \{n_{\min\kappa}\}$$

Ein Beispiel für den Vergleich einer Cauer-Realisierung mit einer Approximation nach Tschebyscheff im Durchlaßbereich wird im Abschnitt 3.4 gegeben.

3.3.2 Die Auswirkung der Filtergraderweiterung auf den Realisierungsaufwand

Für eine große Anzahl unterschiedlicher Toleranzschemata wurden Cauer-Filter unterschiedlichen Grades berechnet und deren Koeffizientenwortlänge mit dem in Abschnitt 3.2 beschriebenen Programm optimiert. Die Ergebnisse der Aufwandsabschätzung nach Gl. (3.23) sind in einem Diagramm, Bild 3.10, in Abhängigkeit von der relativen Graderhöhung $\Delta n/n_{\min}$ eingetragen.

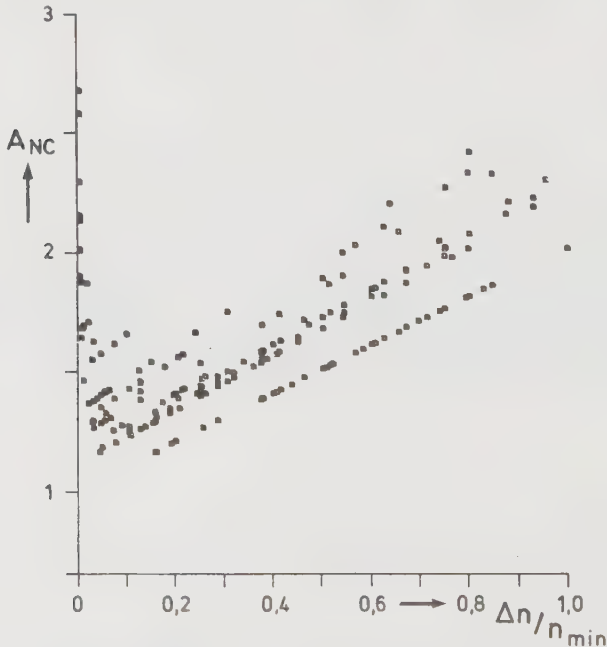


Bild 3.10: Aufwandsmaß für unterschiedliche Werte $\Delta n/n_{\min}$

Man sieht, wie der Aufwand einerseits für $\Delta n/n_{\min} \rightarrow 0$ steil ansteigt, andererseits ungefähr linear mit der Graderweiterung für große Werte zunimmt.

Ein ausgeprägtes Minimum nimmt die Aufwandsabschätzung für $\Delta n/n_{\min} \approx 0.1$ an. Dies bedeutet, daß man bei einer Cauer-

Realisierung und einer Graderweiterung von ungefähr 10 % eine optimale Lösung erwarten kann.

Entsprechende Untersuchungen wurden für Potenzfilter durchgeführt. Der minimale Aufwand liegt hier bei einer Graderweiterung von ungefähr 5 %.

Zur Bestimmung einer automatischen Graderhöhung im Rahmen des Entwurfsprogramms wird n_{opt} zweckmäßig als

$$n_{\text{opt}} = \max \left\{ \begin{array}{l} \left[n_{\min} \left(1 + \frac{\Delta n}{n_{\min}} \right)_{\text{opt}} \right] \text{ gerundet} \\ \left[n_{\min} \right] \text{ aufgerundet} \end{array} \right\} \quad (3.25)$$

gewählt.

Darüber hinaus kann es sinnvoll sein, aus organisatorischen Gründen den nächsten geraden Filtergrad zu wählen, um stets mit Blöcken zweiten Grades arbeiten zu können.

3.4 Beispiele zur Koeffizientenoptimierung

In den vorausgehenden Abschnitten wurde ein Verfahren zur Koeffizientenoptimierung durch geeignete Wahl der Entwurfskonstanten C vorgeschlagen. Ein allgemeines Aufwandsmaß wurde zur Beurteilung der verschiedenen Entwurfsmöglichkeiten angegeben. Anhand von einigen Beispielen werden Ergebnisse gezeigt, die mit dem Programmsystem erzielt wurden.

3.4.1 Cauer-Tiefpässe mit unterschiedlicher Filtergrad- erweiterung

Im Abschnitt 2.2.2 wurde im Bild 2.5 das Toleranzschema eines Tiefpasses angegeben. Nach Tabelle 2.2 benötigt man für eine

Cauer-Approximation einen minimalen Filtergrad $n_{\min} = 7.99$. Das Toleranzschema wird also gerade mit einem Tiefpaß 8. Grades erfüllt (siehe auch Abschnitt 3.1.1). Dabei benötigt man bei einer gleichmäßigen Ausnutzung des Toleranzbereiches bei ungerundeten Koeffizienten eine notwendige Wortlänge $w_c = 19$ Bit. Der Frequenzgang dieses Filters wird in Bild 2.6c gezeigt. Der entsprechende Datensatz ist dem Anhang (Beispiel 1) zu entnehmen. Durch eine Optimierung der Entwurfskonstanten C kann die notwendige Koeffizientenwortlänge auf $w_c = 16$ Bit reduziert werden (Datensatz: Anhang, Beispiel 2). Eine weitere Möglichkeit zur Reduzierung der Koeffizientenwortlänge ergibt sich durch die Filtergraderweiterung. Die Ergebnisse für Filter bis zum Grad 12 werden zusammen mit dem Aufwand nach Gl. (3.21) in Tabelle 3.1 zusammengestellt.

Beispiel Nr.	Filtergrad n	$\Delta n/n_{\min}$	w_c	A_{NC}
2	8	≈ 0.001	16	2.29
3	9	0.125	9	1.45
4	10	0.25	7	1.25
6	11	0.377	7	1.38
7	12	0.50	7	1.50

Cauer-Tiefpaß: $\Omega_D = \pi/3$, $\Omega_S = 1.18$, $\delta_D = 0.02$, $\delta_S = 0.001$
 $w_{c\min} = 7$

Tabelle 3.1: Filtergraderweiterung beim Cauer-Tiefpaß.

Wie man sieht, ergibt sich hier eine im Bezug auf die Aufwandszahl optimale Lösung für einen Filtergrad $n = 10$. Die stark unterschiedliche Grundaussnutzung des Toleranzschemas wird deutlich, wenn man in Bild 3.11 die beiden Frequenzgänge für gerundete und ungerundete Koeffizienten vergleicht.

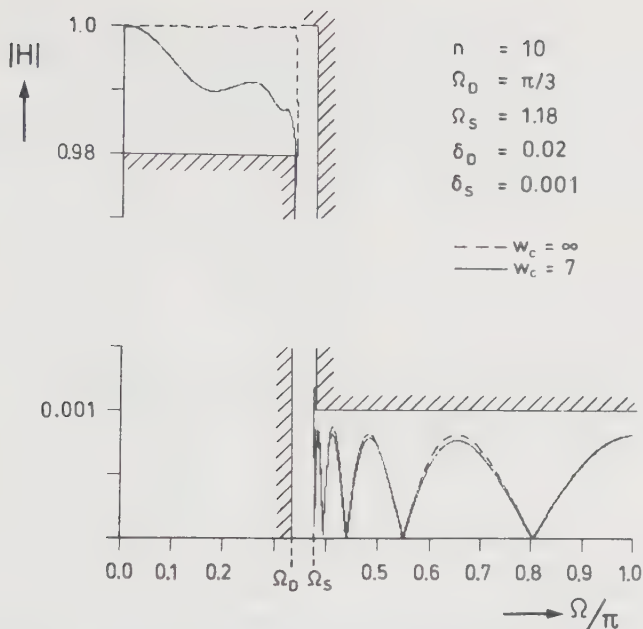


Bild 3.11: Frequenzgänge des Beispiels Nr. 4

3.4.2 Tschebyscheff-Tiefpässe mit unterschiedlicher Filtergraderweiterung

Für das Toleranzschema nach Bild 2.5 werden zum Aufwandsvergleich neben dem in Bild 2.6 dargestellten Beispiel Nr. 8 weiter Filter mit Tschebyscheff-Verhalten im Durchlaßbereich berechnet. Wie aus der Tabelle 3.2 zu entnehmen ist, kann auch hier durch eine Filtergraderweiterung eine Reduzierung der notwendigen Koeffizientenwortlänge erzielt werden. Die günstigste Lösung ergibt sich für einen Filtergrad $n = 18$. Das Beispiel zeigt, daß eine Filtergradserweiterung $\Delta n/n_{\min} = 0.05 - 0.1$ auch für Approximationen mit Tschebyscheff-Verhalten im Durchlaßbereich geeignet ist.

Beispiel Nr.	Filtergrad n	$\Delta n/n_{\min}$	w_c	A_{NC}
9	17	0.032	12	1.24
10	18	0.093	11	1.20
*	19	0.153	11	1.27
*	20	0.214	11	1.33
*	21	0.275	10	1.27
*	22	0.335	10	1.33

Tabelle 3.2: Tschebyscheff-Tiefpaß: $\Omega_D = \pi/3$, $\Omega_S = 1.18$,
 $\delta_D = 0.02$, $\delta_S = 0.001$
 $w_{c_{\min}} = 10$

* Datensatz nicht im Anhang aufgeführt.

Ein Vergleich der Cauer-Lösungen nach Abschnitt 3.4.1 mit den hier berechneten Tschebyscheff-Approximationen ist anhand des allgemeinen Bewertungsmaßes nach Gl. (3.17) möglich. Berücksichtigt man die Einsparung eines Multiplizierers je Block 2. Grades bei der Tschebyscheff-Lösung, so ergibt sich hier dennoch für das Cauer-Filter das günstigere Aufwandsmaß (Tab. 3.3).

	Cauer	Tschebyscheff
$m_{R_K}^*$	3	2
$n_{\min_K}$	7.99	16.47
$w_{c_{\min_K}}$	7	10
A_{R_K}	1.5	2.944
günstigste Lösung		
n	10	18
A_{NC}	1.25	1.20
A	2.50	3.53

Tabelle 3.3: Vergleich der Cauer- und Tschebyscheff-Lösungen

* ohne Berücksichtigung der Skalierung

3.4.3 Verschiedene Approximationen der Bandsperre nach Golden und Kaiser

Von Golden und Kaiser wurde in [54] ein Toleranzschema für eine hochselektive Bandsperre mit folgenden Daten angegeben:

Abtastfrequenz 10 kHz

$$f_{S1} = 2.596; \quad f_{S2} = 2.836 \text{ kHz} \quad \hat{=} \quad \Omega_{S1} = 1.631, \\ \Omega_{S2} = 1.782$$

$$f_{D1} = 2.588; \quad f_{D2} = 2.844 \text{ kHz} \quad \hat{=} \quad \Omega_{D1} = 1.626, \\ \Omega_{D2} = 1.767$$

$$a_{\max} = -0.5 \text{ dB} \quad \hat{=} \quad \delta_D = 0.05594$$

$$a_{\min} = -75 \text{ dB} \quad \hat{=} \quad \delta_S = 0.00018$$

Der Entwurf erfolgte dabei so, daß der Durchlaßbereich voll ausgenützt wurde. Aufgrund des ganzzahligen Filtergrades $n=22$ ergibt sich somit eine Sperrdämpfung von -76.5 dB .

Wie bereits die Untersuchungen in [6] gezeigt haben, ist bei der Bandsperre 22. Grades die Koeffizientenempfindlichkeit ebenfalls im Durchlaßbereich größer als im Sperrbereich. Durch eine Optimierung der Entwurfskonstanten ist auch hier eine Verbesserung möglich. In Tabelle 3.4 werden Approximationen des vorgegebenen Toleranzschemas mit unterschiedlicher Filtergrad-erhöhung verglichen. Die in [54] vorgeschlagene Lösung wird der Tabelle vorangestellt.

Auch ohne Filtergraderweiterung erbringt die Optimierung eine starke Reduzierung der notwendigen Koeffizientenwortlänge. Eine geringe Verbesserung kann man darüber hinaus noch durch eine Graderhöhung auf $n = 24$ erreichen. Einen Vergleich der Frequenzgänge der Beispiele 14 und 16 mit gerundeten Koeffizienten zeigt das Bild 3.12.

Beispiel Nr.	Filtergrad n	$\Delta n/n_{\min}$	w_c	A_{NC}	A
14 (nach [54])	22	0.0157	20	1.693	2.032
15	22	0.0157	16	1.354	1.625
16	24	0.1081	14	1.293	1.552
*	26	0.2004	13	1.300	1.560
*	28	0.2928	13	1.400	1.680
*	30	0.3851	12	1.385	1.662

Tabelle 3.4: Aufwandsvergleich der Bandsperre nach Golden und Kaiser [54]

$$n_{\min} = 21.658 \quad w_{c_{\min}} = 12$$

* Datensatz nicht im Anhang aufgeführt

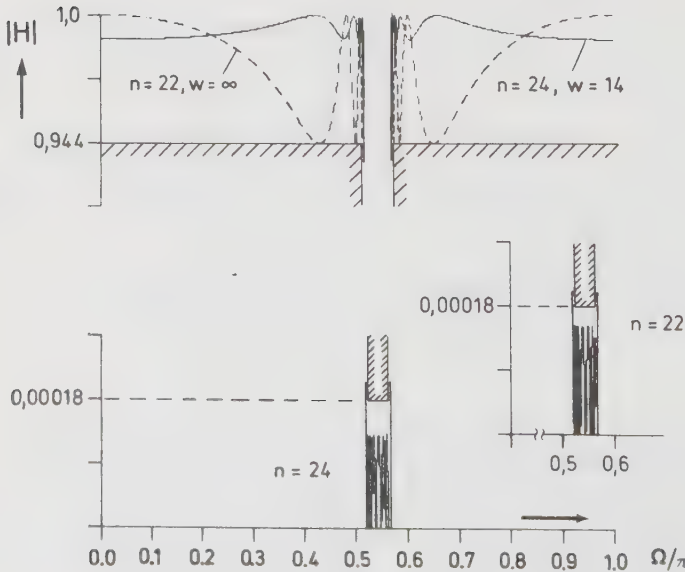


Bild 3.12: Frequenzgang der Bandsperre 22. Grades ([54]) und 24. Grades

Eine weitere Verkürzung der Koeffizientenwortlänge kann durch die Differenzdarstellung der Koeffizienten c_o und eine für jeden Koeffizienten getrennte Pseudo-Gleitkomma-Darstellung (Abschnitt 2.3.1) erreicht werden.

Die Koeffizienten c_{ov} werden dabei durch die Differenz-Koeffizienten c'_{ov} dargestellt

$$c'_{ov} = 1 - c_{ov}$$

Die freie Wahl der Exponenten wird auf $e > -4$ beschränkt. Da der Exponent hier maximal $e = -1$ (siehe Datensatz) werden kann, benötigt man für eine zusätzliche Abspeicherung $i_E = 2$ Bit. Für die Bewertung der Schaltung soll jedoch hier nur die der Mantisse entsprechende Multiplizierer-Wortlänge gerechnet werden. Einen Überblick über die notwendigen Wortlängen und den Aufwand gibt die Tabelle 3.5.

Beispiel Nr.	Filtergrad n	$\Delta n/n_{\min}$	w_c (ohne Exp.)	$A_{NC}=A$
*	22	0.0157	13	1.320
17	24	0.1081	11	1.219
*	26	0.2004	10	1.200
*	28	0.2928	10	1.293

Tabelle 3.5: Aufwandsvergleich der Bandsperre nach Golden und Kaiser bei Pseudo-Gleitkomma- und Differenz-Darstellung der Koeffizienten

* Datensatz nicht im Anhang aufgeführt

Vergleicht man die notwendigen Multiplizierer-Wortlängen der beiden Realisierungen, so können 3 Bit bei der zweiten Art eingespart werden. Durch Bewertung des Aufwandsmaßes A_{NC} mit

$$A_R = \frac{w_{cmin1}}{w_{cmin2}} = \frac{12}{10} = 1.2 \text{ wurde in Tabelle 3.4 das vergleichbare}$$

Aufwandsmaß A zusätzlich berechnet.

Der Vergleich der jeweils günstigsten Lösungen führt zu einer Aufwandsreduzierung um etwa 29 %.

3.4.4 Tiefpaß 10. Grades mit Skalierung

Werden die zur Minimisierung des Rauschens am Ausgang des Filters notwendigen Skalierungsfaktoren direkt in die Zählerkoeffizienten mit einbezogen, so wird eine mehrstufige Optimierung (Abschnitt 3.2.3 und 5.3.3) notwendig. Ein Tiefpaß 10. Grades, der das in Bild 2.5 gezeigte Toleranzschema erfüllt, wurde unter diesen Bedingungen entworfen (Beispiel 5). Die notwendige Koeffizientenwortlänge liegt bei dem Ergebnis um ein Bit höher als beim entsprechenden Filter mit Vorfaktoren. Der Frequenzgang mit gerundeten und ungerundeten Koeffizienten ist in Bild 3.13 dargestellt.

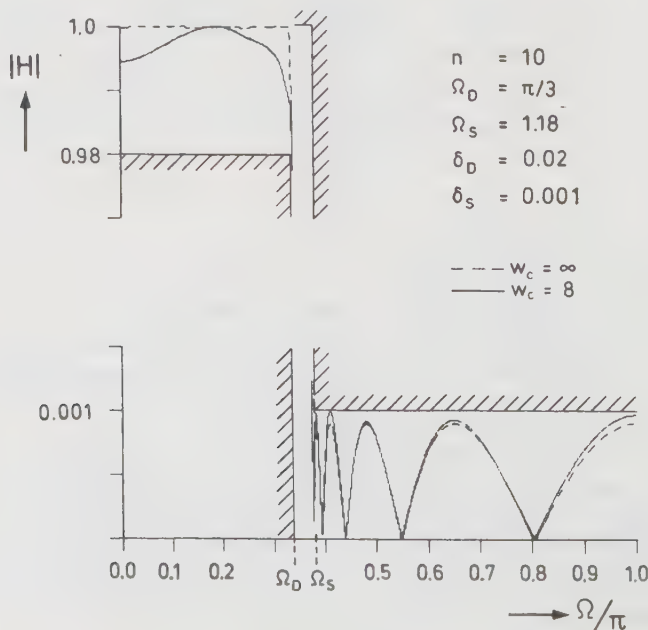


Bild 3.13: Frequenzgang für Tiefpaß 10. Grades mit Skalierung.

4. Theoretische Grundlagen für die Bewertung des Rauschverhaltens verschiedener Strukturen mit unterschiedlicher Arithmetik

Die Veränderung der Übertragungsfunktion eines digitalen Systems aufgrund der endlichen Koeffizientenwortlänge wurde in den vorausgehenden Abschnitten behandelt und eine Optimierung dieses sogenannten "Einstellfehlers" durchgeführt. Die endliche Länge der Zustandsvariablen und somit die begrenzte Genauigkeit der mathematischen Operationen führt zu den sogenannten "Betriebsfehlern" der digitalen Filter. Dabei sind die rauschartigen Fehler beim normalen Betrieb des Filters und die Grenzyklen beim Übergang auf ein konstantes Eingangssignal zu unterscheiden.

Die in den folgenden Abschnitten durchgeführten Untersuchungen beschäftigen sich mit den rauschartigen Fehlern bei verschiedenen Strukturen mit unterschiedlicher Rundungs-Arithmetik. Die aus der Literatur [1] - [4], [7], [55] - [58] bekannten und für die Bestimmung der Aufwandsmaße notwendigen Fehlermodelle werden dargestellt und zur Anwendung auf spezielle Rundungs-Arithmetiken erweitert.

Bekanntlich läßt sich eine gegebene Übertragungsfunktion nach Gl. 2.6

$$H(z) = \frac{\sum_{\mu=0}^m b_{\mu} z^{\mu}}{\sum_{v=0}^n c_v z^v} ; \quad c_n = 1$$

mit unterschiedlichen Strukturen realisieren, die im allgemeinen auch unterschiedliche Rauschleistungen am Ausgang aufweisen. Zur Bewertung der einzelnen Strukturen werden geeignete Aufwandsmaße angegeben, die im Gegensatz zur ausschließlichen Betrachtung der Rauschleistung eine direkte Beurteilung des schaltungstechnischen Aufwandes ermöglichen.

Eine außerordentlich große Zahl von Strukturen wurde bisher

vorgeschlagen. Da sich diese Strukturen auch noch durch die Zahlendarstellung und die Art der verwendeten Arithmetik variieren lassen, ist eine Beschränkung der Untersuchungen auf bestimmte Gruppen von Realisierungen notwendig.

4.1 Zahlendarstellung der Zustandsvariablen

In [3], [59] wird die Gleitkomma-Darstellung der Koeffizienten und Zustandsvariablen erläutert und deren Rauschverhalten diskutiert. Die Realisierung eines digitalen Filters zur Sprachverarbeitung mit Gleitkomma-Arithmetik wird in [60] dargestellt und dessen günstige Eigenschaften in Bezug auf den Dynamik-Bereich und die Reduzierung von Grenzyklen gezeigt.

Eine Mischung von Gleitkomma- und Festkomma-Arithmetik stellt die Block-Gleitkomma-Arithmetik [3] dar, bei der nur am Eingang und Ausgang der einzelnen Schaltungsblöcke eine Umnormierung stattfindet. Die Anwendung dieser Arithmetik erscheint nach [3] besonders günstig für den Aufbau von Fourier-Prozessoren. In gewisser Weise kann auch eine Kaskadenschaltung von Teilblöcken 2. Grades mit Vorfaktoren in Zweier-Potenzen (Abschnitt 5.3.4.1) als eine Art Block-Gleitkomma-Darstellung betrachtet werden.

Eine echte Gleitkomma-Darstellung der Zustandsvariablen benötigt im Vergleich zur Festkomma-Darstellung einen größeren schaltungstechnischen Aufwand. Darüber hinaus stören die notwendigen Normalisierungen den Ablauf einer seriellen Verarbeitung. Die folgenden Ausführungen bleiben daher auf die Festkomma-Arithmetik mit Dualzahlen beschränkt.

Die Darstellung der negativen Zahlen wird dabei im Zweier-Komplement durchgeführt.

$$x = -\xi_{i_2+1} 2^{i_2+1} + \sum_{i=i_1}^{i_2} \xi_i 2^i ; \quad \xi_i = 0, 1; \quad i = i_1(1) i_2 \quad (4.1)$$

Man erhält eine endliche Anzahl von $2^{i_2 - i_1 + 2}$ unterschiedlichen Zahlenwerten, die in einem Bereich

$$-2^{i_2 + 1} \leq x < 2^{i_2 + 1} \quad (4.2)$$

liegen können. Die Quantisierungsstufe dieser Zahlen beträgt $Q = 2^{i_1}$. Ohne Verlust an Allgemeinheit wird $i_2 = -1$ gewählt und somit die Aussteuerungsgrenze der Festkommazahlen auf

$$-1 \leq x < 1 \quad (4.3)$$

festgelegt. Bezeichnet man mit $q = -i_1$ die Anzahl der Stellen nach dem Komma, so ergibt sich als notwendige Gesamtlänge

$$w = q + 1 = 1 - i_1 \quad (4.4)$$

und die Darstellung nach Gl. (4.1) geht über in

$$x = -\xi'_w + \sum_{i=1}^{w-1} \xi'_i 2^{i-q-1} \quad \xi'_i \in \{0, 1\} \quad (4.5)$$

Hier sind 2^w unterschiedliche Zahlen möglich. Die Quantisierung beträgt

$$Q = 2^{-q} \quad (4.6)$$

4.2 Modelle zur Berechnung des Quantisierungsrauschens bei unterschiedlichen Rundungsarithmetiken

Die Darstellung der analogen Signalgrößen durch digitale Zahlenwerte mit endlicher Genauigkeit verursacht bei der digitalen Signalverarbeitung Quantisierungsfehler, die im folgenden anhand von statistischen Modellen näher untersucht werden sollen.

Bei den Betrachtungen ist prinzipiell zwischen einer kontinuierlichen Fehlerfolge, wie sie bei der Umwandlung eines zunächst kontinuierlichen Signals $u(t)$ in die Folge diskreter Werte endlicher Wortlänge $\{[u(k)]_Q\}$ im Analog-Digital-Wandler entsteht und einer diskreten Fehlerfolge zu unterscheiden, wie sie bei der Verkürzung bereits digital vorliegender Signale auftritt. Die unterschiedlichen Verteilungen der Fehler wurden von Eckhardt in [7] ausführlich untersucht. Im ersten Fall erhält man eine kontinuierliche, im zweiten eine diskrete Verteilung. Wird jedoch bei der Verkürzung digitaler Signale eine hinreichend große Quantisierung im Vergleich zum Ausgangssignal vorgenommen,

$$2^{-q_1} \ll 2^{-q_s}$$

mit q_1 = Wortlänge des unverkürzten Wertes

q_s = Wortlänge des verkürzten Wertes nach dem Festkomma

so kann auch im zweiten Fall die Fehlerverteilung in guter Näherung als kontinuierlich angesehen werden, so daß sich die entsprechenden Modelle verwenden lassen [7].

4.2.1 Modelle mit kontinuierlicher Fehlerverteilung

Die verschiedenen Verkürzungsarithmetiken werden am Beispiel der Analog-Digital-Wandlung erklärt und deren unterschiedliche Fehleranteile bestimmt.

Entsprechend dem Bild 4.1 erfolgt die Aufteilung des A/D-Wandlers in einem ideal arbeitenden Teil, der das Eingangssignal $u(t)$ in eine diskrete Folge $\{u(k)\}$ beliebiger Genauigkeit umsetzt und in einen Quantisierer, der diese Folge auf die vorgegebene Wortlänge $w = w_s$ verkürzt. Es entsteht dabei ein Fehler

$$\Delta(k) = [u(k)]_Q - u(k) , \quad (4.7)$$

wenn mit $[u(k)]_Q$ das quantisierte Eingangssignal bezeichnet wird. Allgemein läßt sich der Fehler $\Delta(k)$ in drei Teile aufspalten, einen rauschartigen Anteil $\Delta^{(R)}(k)$, einen mit dem Signal korrelierten $\Delta^{(S)}(k)$ und einen Gleichanteil $\Delta^{(=)}(k)$.

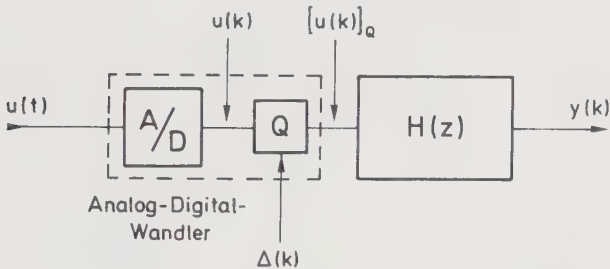


Bild 4.1: Analog-Digital-Wandlung mit endlicher Wortlänge

$$\Delta(k) = \Delta^{(R)}(k) + \Delta^{(S)}(k) + \Delta^{(=)}(k) \quad (4.8)$$

Durch Mitteln der Fehlerquadrate kann die Fehlerleistung am Ausgang des Quantisierers berechnet werden.

$$N = \lim_{m \rightarrow \infty} \frac{1}{2m+1} \sum_{u=-m}^{+m} \Delta^2(k) = \overline{\Delta^2(k)} \quad (4.9)$$

Sind die einzelnen Fehleranteile in Gl. (4.8) miteinander unkorreliert, so läßt sich die mittlere Fehlerleistung entsprechend aufteilen

$$N = N^{(R)} + N^{(S)} + N^{(=)} \quad (4.10)$$

4.2.1.1 Runden

Führt man die Verkürzung der Zustandsvariablen entsprechend der Rundungsvorschrift

$$[x]_R = \text{sgn}(x) \cdot Q \cdot [Q^{-1} |x| + 0.5] \quad \begin{matrix} 1) \\ 2) \end{matrix} \quad (4.11)$$

durch, so ergibt sich der Fehler zu

$$\Delta_R = [x]_R - x \quad (4.12)$$

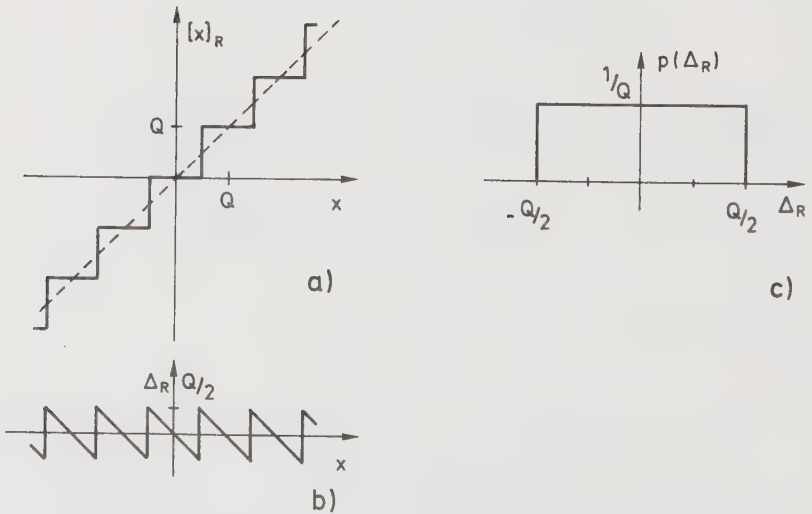


Bild 4.2: Quantisierungskennlinien und Fehlerverteilung bei Runden

1) $[\cdot]$ so, daß $[x]$ größte ganze Zahl mit $x - [x] \geq 0$

2) $[\cdot]_Q, [\cdot]_R, [\cdot]_T \dots$ quantisierte Größe allgemein, bzw. nach entsprechender Vorschrift.

Die Quantisierungskennlinie für Signal und Fehler zeigt Bild 4.2a,b.

Mit den Annahmen

- die Fehlerfolge $\{\Delta(k)\}$ ist eine Abtastfolge eines stationären weißen Zufallsprozesses
- innerhalb jeder Quantisierungsstufe sind alle Werte gleich wahrscheinlich
- die Fehlerfolge ist mit dem exakten Signal $\{u(k)\}$ unkorreliert ($\Delta^{(S)}(k) = 0$)

und der Quantisierung Q nach Gl. (4.6) erhält man die mittlere Fehlerleistung

$$\overline{\Delta_R^2} = \frac{1}{Q} \int_{-Q/2}^{+Q/2} \Delta_R^2 d\Delta_R = \frac{Q^2}{12} \quad (4.13)$$

Entsprechend der Verteilung nach Bild 4.2c ist die Fehlerfolge $\Delta_R(k)$ gleichwertfrei,

$$\overline{\Delta_R} = e_R^{(=)} = \frac{1}{Q} \int_{-Q/2}^{Q/2} \Delta_R d\Delta_R = 0 \quad 1) \quad (4.14)$$

Die Varianz σ_R^2 der Ersatzrauschquelle beträgt somit

$$\sigma_R^2 = \sigma_R^{2(R)} = \overline{\Delta_R^2} - \overline{\Delta_R}^2 = \frac{Q^2}{12} \quad 2) \quad (4.15)$$

Entsprechend Bild 4.1 wird der vom Quantisierer hervorgerufene Fehler durch das nachfolgende System übertragen und erfährt dabei eine Bewertung entsprechend der Übertragungsfunktion des Systems $|1|$. Die im Innern eines Filtersystems notwendigen

1) $e^{(=)}$ Gleichwertanteil bei beliebiger Fehlerquelle

2) $\sigma^{2(R)}$ = Varianz des rauschartigen Störanteils einer beliebigen Fehlerquelle

Quantisierungen können bei Runden in gleicher Weise behandelt werden. Ihre Anteile zum Gesamtrauschen summieren sich am Ausgang (siehe Abschnitt 4.2.3).

In Gl. (4.13) wurde die mittlere Fehlerleistung unter der Annahme eines rauschförmigen Eingangssignals bestimmt. Knowles und Edwards haben in [55] gezeigt, daß das zugrunde liegende statistische Modell unter bestimmten Voraussetzungen (siehe 4.2.3.2b) auch dann angewandt werden kann, wenn das zu quantisierende Signal sinusförmig ist.

4.2.1.2 Zweierkomplement-Abschneiden

Die einfachste Art einer Wortlängenverkürzung bei der Zahlendarstellung im Zweierkomplement ist das Abschneiden der Bits hinter der gewünschten Quantisierungsstufe.

$$[x]_{TT} = Q [Q^{-1} x] \quad (4.16)$$

Wie aus der Quantisierungskennlinie in Bild 4.3 (TT = Two's Complement Truncation) zu entnehmen ist, entsteht bei dem 2er-Komplement-Abschneiden stets eine Verminderung des Zahlenwerts.

Für die mittlere Fehlerleistung $\overline{\Delta_{TT}^2}$ erhält man nach [7]

$$\overline{\Delta_{TT}^2} = \frac{1}{Q} \int_{-Q}^0 \Delta_{TT}^2 d\Delta_{TT} = \frac{Q^2}{3}, \quad (4.17)$$

für den linearen Mittelwert des Fehlers $\overline{\Delta_{TT}}$

$$\overline{\Delta_{TT}} = e_{TT}^{(=)} = \frac{1}{Q} \int_{-Q}^0 \Delta_{TT} d\Delta_{TT} = -\frac{Q}{2}, \quad (4.18)$$

und für die Varianz σ_{TT}^2 des rauschartigen Fehleranteils

$$\sigma_{TT}^2 = \sigma_{TT}^2(R) = \overline{\Delta_{TT}^2} - \overline{\Delta_{TT}}^2 = \frac{Q^2}{3} - \frac{Q^2}{4} = \frac{Q^2}{12} \quad (4.19)$$

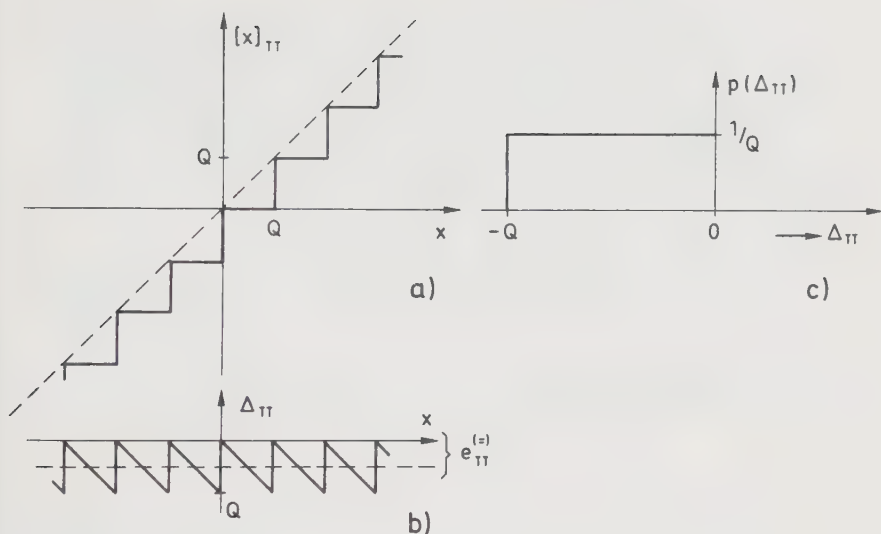


Bild 4.3: Quantisierungskennlinie und Fehlerverteilung
bei 2er-Komplement-Abschneiden

Ein Vergleich mit Gl. (4.15) zeigt, daß die Varianzen für beide Arithmetiken übereinstimmen, die mittlere Leistung der Fehlerquelle bei Zweierkomplement-Abschneiden durch den zusätzlichen Gleichanteil jedoch größer ist.

Bei der Berechnung der Rauschleistung mehrerer Nichtlinearitäten mit Zweierkomplement-Abschneiden summieren sich die rauschartigen Leistungsanteile am Ausgang. Die Gleichanteile summieren sich direkt (siehe Abschnitt 4.2.3).

4.2.1.3 Betragsmäßiges Abschneiden

Für positive Zahlen entspricht das betragsmäßige Abschneiden ($T \hat{=}$ Truncation; in $|7|$ mit betragsmäßiges Abrunden bezeichnet) dem Zweierkomplement-Abschneiden; der Wert einer Zahl wird betragsmäßig verringert. Unterschiede ergeben sich für negative Zahlen. Hier bedeutet die betragsmäßige Verringerung der negativen Zahlen eine Erhöhung des Zahlenwerts.

$$[x]_T = \text{sgn}(x) \cdot Q \cdot [Q^{-1}|x|] \quad (4.20)$$

Die Quantisierungskennlinie für betragsmäßiges Abschneiden wird in Bild 4.4a gezeigt.

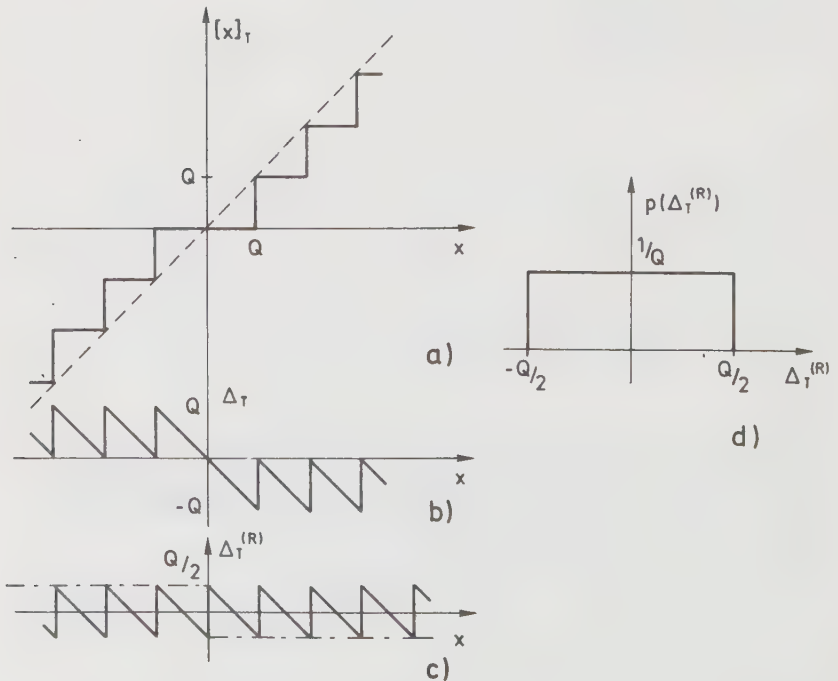


Bild 4.4: Quantisierungskennlinie, Fehlerkennlinie und Fehlerverteilung bei betragsmäßigem Abschneiden

Anhand der Fehlerkennlinie (Bild 4.4b) wird die Abhängigkeit des Fehlers vom Vorzeichen des Signals deutlich. Nach [7], [58] kann der Fehler bei betragsmäßigem Abschneiden in einen mit dem Signal korrelierten Anteil $-\frac{Q}{2} \operatorname{sgn}(x)$ und einen unkorrelierten rauschartigen Anteil $\Delta_T^{(R)}$ aufgeteilt werden:

$$\Delta_T(k) = -\frac{Q}{2} \operatorname{sgn}(x(k)) + \Delta_T^{(R)}(k) \quad (4.21)$$

Wie man im Bild 4.4c anhand der Fehlerkennlinie sieht, entspricht diese bis auf eine Verschiebung [58] der Kennlinie bei Runden.

Die mittlere Fehlerleistung $\overline{\Delta_T^2}$ errechnet sich nach [7] zu

$$\overline{\Delta_T^2} = \frac{1}{2Q} \int_{-Q}^Q \Delta_T^2 d\Delta_T = \frac{Q^2}{3} \quad (4.22)$$

Aufgrund der symmetrischen Fehlerkennlinie verschwindet der Gleichanteil $e_T^{(=)}$

$$\overline{\Delta_T} = e_T^{(=)} = \frac{1}{2Q} \int_{-Q}^Q \Delta_T d\Delta_T = 0 \quad (4.23)$$

Neben den bereits eingeführten Kennwerten wird hier die Amplitude des mit dem Signal korrelierten Fehleranteils $e_T^{(S)}$ eingeführt,

$$e_T^{(S)} = \frac{1}{Q} \int_0^Q \Delta_T d\Delta_T = -\frac{Q}{2} \quad (4.24)$$

Die Varianz des rauschartigen Anteils errechnet sich somit zu

$$\sigma_T^{2(R)} = \overline{\Delta_T^2} - (e_T^{(S)})^2 = \frac{Q^2}{3} - \frac{Q^2}{4} = \frac{Q^2}{12} \quad (4.25)$$

Da ein Teil des Fehlersignals mit dem Signal korreliert ist, geht bei der Bestimmung der Fehlerleistung N am Ausgang eines Filtersystems neben der Übertragungsfunktion von der Nichtlinearität zum Ausgang $G(z)$ auch die spektrale Leistungsdichte

$\phi_{xx}(z)$ des Signals am Eingang der Nichtlinearität in die Berechnung ein [7], [58]. Bei einem weißen Rauschen am Eingang ist die Leistungsdichte $\phi_{xx}(z)$ eine Funktion der Übertragungsfunktion $F(z)$ vom Eingang zur Nichtlinearität (Bild 4.5). Es ergibt sich eine gefärbte spektrale Leistungsdichte des Fehlers $\phi_{\Delta T \Delta T}(z)$ und somit die Leistung

$$N = N_T = f(Q, G(z), F(z)) \quad (4.26)$$

Eine allgemeine Berechnung erfolgt im Abschnitt 4.2.3.

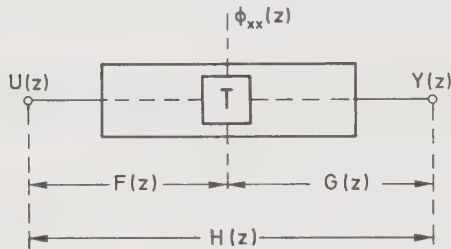


Bild 4.5: Übertragungssystem mit einer Abschneidestelle

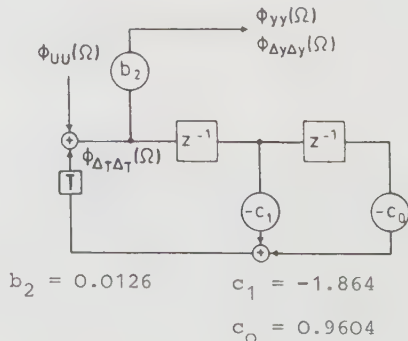
Der Einfluß der spektralen Leistungsdichte des Eingangssignals auf die Leistungsdichte des Fehlers soll im folgenden anhand von drei Beispielen erläutert werden. Die Berechnung der dargestellten Kurven wird dabei entsprechend dem allgemeinen Ansatz nach Abschnitt 4.2.3 durchgeführt. Die dort eingeführten Bezeichnungen werden hier verwendet.

Als Übertragungssystem wird ein rein rekursives System 2. Grades mit einer Nichtlinearität gewählt.

$$H(z) = \frac{b_2 z^2}{z^2 + c_1 z + c_0}$$

$$F(z) = \frac{-c_1 z - c_0}{z^2 + c_1 z + c_0}$$

$$G(z) = H(z)$$



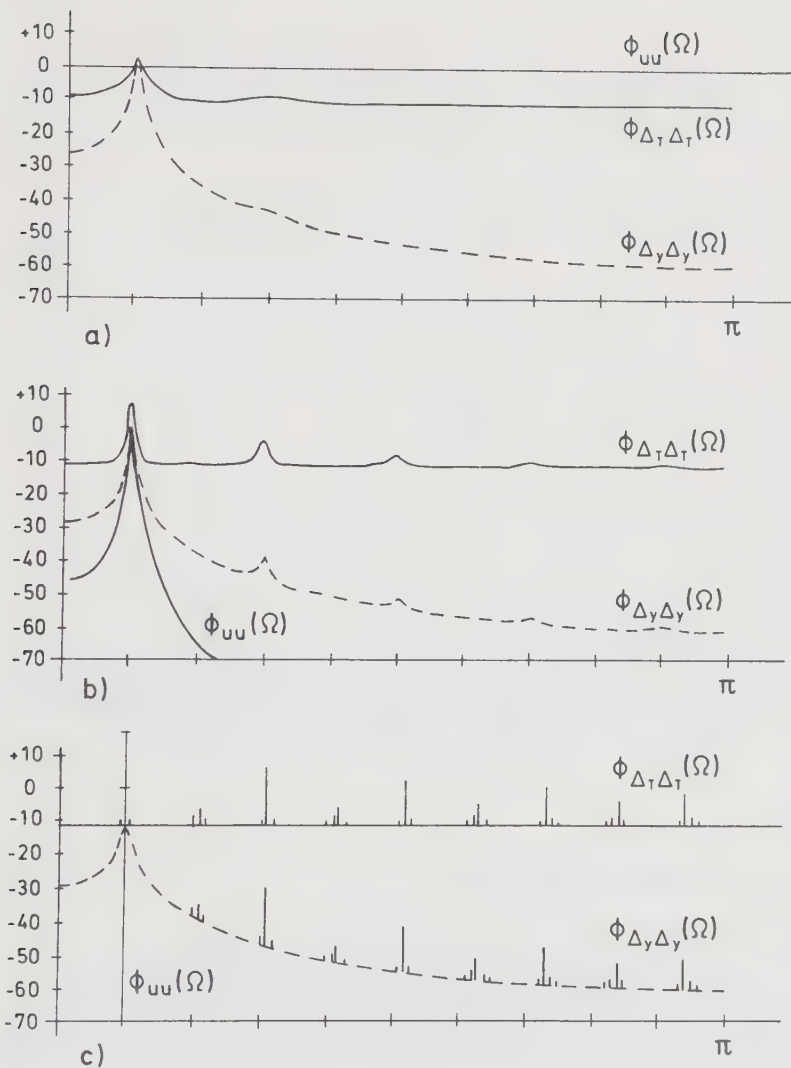


Bild 4.6: Fehlerleistungsspektren bei verschiedenen Eingangssignalen

a) weißes Rauschen

b) farbiges Rauschen

c) sinusförmig

ϕ_{uu} = Signalspektrum

$\phi_{\Delta_T \Delta_T}$ = Fehlerspektrum am Ausgang der Nichtlinearität

$\phi_{\Delta_y \Delta_y}$ = Fehlerspektrum am Filterausgang

Als Eingangssignal wird verwendet:

a) weißes Rauschen $\phi_{uu}(\Omega) = \text{const.} = 1$.

b) farbiges Rauschen $\phi_{uu}(\Omega) = \phi_{uu}(z) = H_f(z) \cdot H_f(z^{-1})$

mit

$$H_f(z) = \left(\frac{1}{z^2 - 1.883 z + 0.98} \right)^2 \quad \text{und } z = e^{j\Omega}$$

c) Sinuserregung mit $\Omega_1 = 0.105 \pi \hat{=} 20^\circ$

Die unterschiedlichen Leistungsspektren der Eingangssignale $\phi_{uu}(\Omega)$ sowie des Fehlers am Quantisierausgang $\phi_{\Delta_T \Delta_T}(\Omega)$ und am Filterausgang $\phi_{\Delta_y \Delta_y}(\Omega)$ werden im Bild 4.6 dargestellt. Da das farbige Eingangsspektrum und die Übertragungsfunktion so aufeinander abgestimmt sind, daß die Resonanzfrequenzen übereinstimmen, wird die starke Anhebung der mit dem Signal korrelierten Spektralanteile deutlich. Für sinusförmige Eingangssignale gehen diese Anteile in diskrete Spektrallinien über.

4.2.1.4 Kontrolliertes Runden

Neben den hier behandelten rauschartigen Betriebsfehlern können bei rekursiven Systemen bei konstantem Eingangssignal sowie beim Abschalten des Eingangssignals Grenzzyklen auftreten [1]. Von verschiedenen Autoren [61] - [66] wurden Strukturen mit bestimmten Arithmetiken zur weitgehenden oder gänzlichen Verminderung von Grenzzyklen vorgeschlagen. Butterweck [63], [64] gelang es mit der von ihm vorgeschlagenen kontrollierten Rundung, Grenzzyklen jeder Art zu vermeiden.

Entsprechend den Vorschriften in [63], [64] wird die kontrollierte Rundung stets so durchgeführt, daß dem rekursiven System

auch bei konstantem Eingangssignal "Pseudo"-Energie ¹⁾ entzogen wird. Hierzu wird eine Bezugsvariable $x_R(k)$ (Bild 4.7) definiert und die Rundung entsprechend der Bedingung

$$| [x(k)]_{CR} - x_R(k) | \leq | x(k) - x_R(k) | \quad (4.27)$$

durchgeführt (CR = controlled rounding).

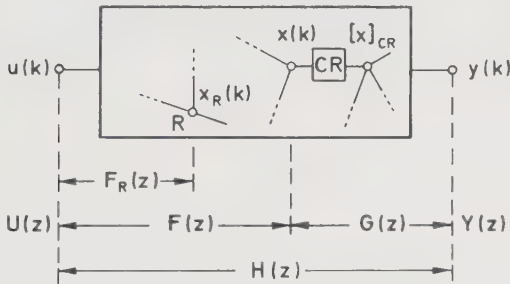


Bild 4.7: System mit kontrollierter Rundung

Führt man die Hilfsgröße

$$x_a(k) = x(k) - x_R(k) \quad (4.28)$$

ein, so erhält man die Verkürzungsvorschrift

$$[x(k)]_{CR} = Q ([Q^{-1} x(k)] + \xi); \quad \xi \in \{0, 1\}$$

$$\begin{aligned} \text{mit} \quad \xi &= 0 & \text{für } x_a(k) \leq 0 \\ \xi &= 1 & \text{für } x_a(k) > 0 \end{aligned} \quad (4.29)$$

Speziell wird in dem in [64] vorgestellten Block 2. Grades die Eingangsgröße als Bezugsgröße verwendet. Für $x_R(k) = 0$ geht das kontrollierte Runden in das betragsmäßige Abschneiden über.

¹⁾ "Pseudo"-Energie, nach [68] Summe der Quadrate der Zustandsvariablen

Entsprechend den Ausführungen für betragsmäßiges Abschneiden läßt sich der Fehler $\Delta_{CR}(k)$ auch hier in einen rauschartigen und einen über die Signum-Funktion mit der Hilfsfunktion $x_a(k)$ korrelierten Anteil aufspalten,

$$\Delta_{CR}(k) = \Delta_{CR}^{(R)}(k) - \frac{Q}{2} \operatorname{sgn}(x_a(k)) . \quad (4.30)$$

Man kann daher die analytische Fehleruntersuchung bei kontrolliertem Runden auf die Untersuchung bei betragsmäßigem Abschneiden zurückführen. Ähnliche Überlegungen wurden hiervon unabhängig von Claasen in [67] durchgeführt.

Durch Transformation der Gl. (4.28) in den Z-Bereich erhält man die Hilfsübertragungsfunktion

$$F_a(z) = F(z) - F_R(z) . \quad (4.31)$$

Dabei bedeutet $F_R(z)$ die Übertragungsfunktion vom Eingang zum Referenzpunkt R (Bild 4.7). Diese Hilfsübertragungsfunktion wird für die weitere Berechnung der Leistungsspektren nach Abschnitt 4.2.3 verwendet. Die Fehlerkennwerte für betragsmäßiges Abschneiden (Abschnitt 4.2.1.3) können für die kontrollierte Rundung übernommen werden.

4.2.2 Modelle mit diskreter Fehlerverteilung

In den vorausgehenden Abschnitten wurde das Rauschverhalten der unterschiedlichen Verkürzungsarithmetiken bei kontinuierlicher Fehlerverteilung untersucht. Unter der Annahme einer Verkürzung um hinreichend viele Stellen ($q_1 \gg q_s$) kann das Modell auch für die weitere Verkürzung bereits quantisiert vorliegender Signale verwendet werden. Eine Multiplikation der Zustandsvariablen mit kurzen Koeffizienten oder eine "shift"-Multiplikation um wenige Stellen anschließende Verkürzung führt jedoch zu einer diskreten Fehlerverteilung [7]. Wie

am Beispiel eines "Shifts" um 1 Bit leicht zu sehen ist, erhält man hier nur die beiden diskreten Fehlerwerte $\Delta = 0$, $Q_S/2$.

Zur Beurteilung von Schaltungen mit "Shift"-Multiplikationen,¹⁾ die besonders für die Skalierung von Teilschaltungen Verwendung finden, werden im folgenden die Fehlermodelle mit diskreter Verteilung bei unterschiedlichen Arithmetiken zusammengestellt. Durch spezielle Rundungsvorschriften für Werte $x = [x]_Q \pm Q/2$ werden die in [7] behandelten Modelle um drei weitere ergänzt.

4.2.2.1 Zweierkomplement Runden

Eine im Zweierkomplement vorliegende Zustandsvariable der Länge $q_1 = q_s + q_c$ nach dem Komma, die z.B. durch Multiplikation einer Zustandsvariablen der Länge q_s mit einem Koeffizienten der Länge q_c entstanden ist, soll wieder auf die Länge q_s gerundet werden. Für die entsprechenden Quantisierungen gilt:

$$\begin{aligned} Q_s &= 2^{-q_s} \\ Q_c &= 2^{-q_c} \\ Q_1 &= 2^{-q_1} \end{aligned} \quad (4.32)$$

Führt man die Verkürzung im Gegensatz zu Gl. (4.11) entsprechend der Vorschrift für Zweierkomplement-Runden ($TR \hat{=} \underline{\text{Two's complement rounding}}$)

$$[x]_{TR} = Q_s \cdot [Q_s^{-1} x + 0.5] \quad (4.33)$$

durch, so erhält man eine unsymmetrische diskrete Quantisierungskennlinie (Bild 4.8). Speziell werden Werte

1) "Shift"-Multiplikation $\hat{=}$ Multiplikation mit Potenzen von zwei

$$x = + (2\nu-1) \cdot Q_S/2 \quad \nu = 1(1)Q_S^{-1}$$

aufgerundet, dagegen die Werte

$$x = - (2\nu-1) \cdot Q_S/2 \quad \nu = 1(1)Q_S^{-1}$$

stets betragsmäßig abgerundet. Das Quantisierungsdiagramm für kurze Koeffizienten wird im Bild 4.8 für $q_C = 1$ und $q_C = 2$ gezeigt.

Aus den möglichen Fehlerwerten wird eine Aufteilung des Gesamtfehlers $\Delta_{TR}(k)$ ersichtlich (Bild 4.8c),

$$\Delta_{TR}(k) = \Delta_{TR}^{(=)}(k) + \Delta_{TR}^{(R)}(k) . \quad (4.34)$$

Mit Hilfe der diskreten Fehlerverteilung läßt sich der Gleichanteil und das mittlere Fehlerquadrat nach [7], [69] berechnen.

$$\overline{\Delta_{TR}} = e_{TR}^{(=)} = \frac{Q_S \cdot Q_C}{2} \quad (4.35)$$

$$\overline{\Delta_{TR}^2} = \frac{Q_S^2}{12} (1 + 2 Q_C^2) \quad (4.36)$$

Da der Fehler keine mit dem Signal korrelierten Anteile enthält,

$$e_{TR}^{(S)} = 0 , \quad (4.37)$$

ergibt sich die Varianz des rauschartigen Fehleranteils zu

$$\sigma_{TR}^2 = \sigma_{TR}^{2(R)} = \overline{\Delta_{TR}^2} - \overline{\Delta_{TR}}^2 = \frac{Q_S^2}{12} (1 - Q_C^2) . \quad (4.38)$$

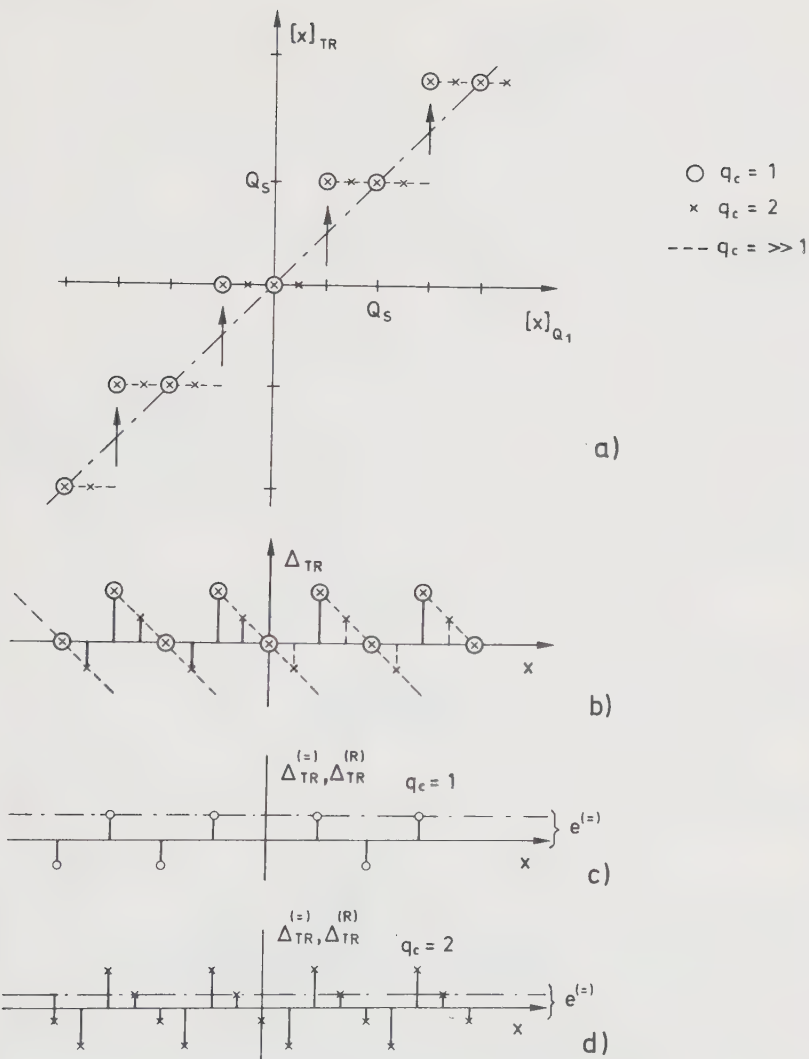


Bild 4.8: Quantisierungsdiagramm und Fehlerwerte bei Zweierkomplement-Runden

4.2.2.2 Betragsmäßiges Runden

Wird jedoch die Quantisierung entsprechend Gl. (4.11)

$$[x]_R = \text{sgn}(x) \cdot Q_R [Q_S^{-1} |x| + 0.5]$$

ausgeführt, so ergibt sich für das betragsmäßige Runden eine symmetrische Kennlinie (Bild 4.9).

Ein Gleichanteil tritt im Fehler nicht auf. Dieser Vorteil wird jedoch mit einem mit dem Signal korrelierten Fehleranteil erkauft. Anhand der im Bild 4.9b gezeigten Fehlerwerte erkennt man, daß sich der Fehler Δ_R in ähnlicher Weise wie beim betragsmäßigen Abschneiden (Abschnitt 4.2.1.3) in einen vom Vorzeichen des Signals abhängigen und einen rauschartigen Anteil aufspalten läßt (Bild 4.9c,d),

$$\Delta_R = \Delta_R^{(S)} + \Delta_R^{(R)} = \text{sgn}(x) e_R^{(S)} + \Delta_R^{(R)} . \quad (4.39)$$

Aufgrund der diskreten Fehlerverteilung erhält man:

$$\overline{\Delta_R} = e_R^{(=)} = 0 , \quad (4.40)$$

mittleres Fehlerquadrat des Gesamtfehlers

$$\overline{\Delta_R^2} = \frac{Q_S^2}{12} (1 + 2 Q_C^2) , \quad (4.41)$$

Kennwert des mit dem Signal korrelierten Fehleranteils

$$e_R^{(S)} = \frac{1}{2} Q_S Q_C ,$$

Varianz des rauschartigen Fehleranteils

$$\sigma_R^{2(R)} = \overline{\Delta_R^2} - (e_R^{(S)})^2 = \frac{Q_S^2}{12} (1 - Q_C^2) . \quad (4.42)$$

Für große q_C geht die Quantisierung $Q_C \rightarrow 0$ und die Kennwerte gehen in die für Runden bei kontinuierlicher Fehlerverteilung über.

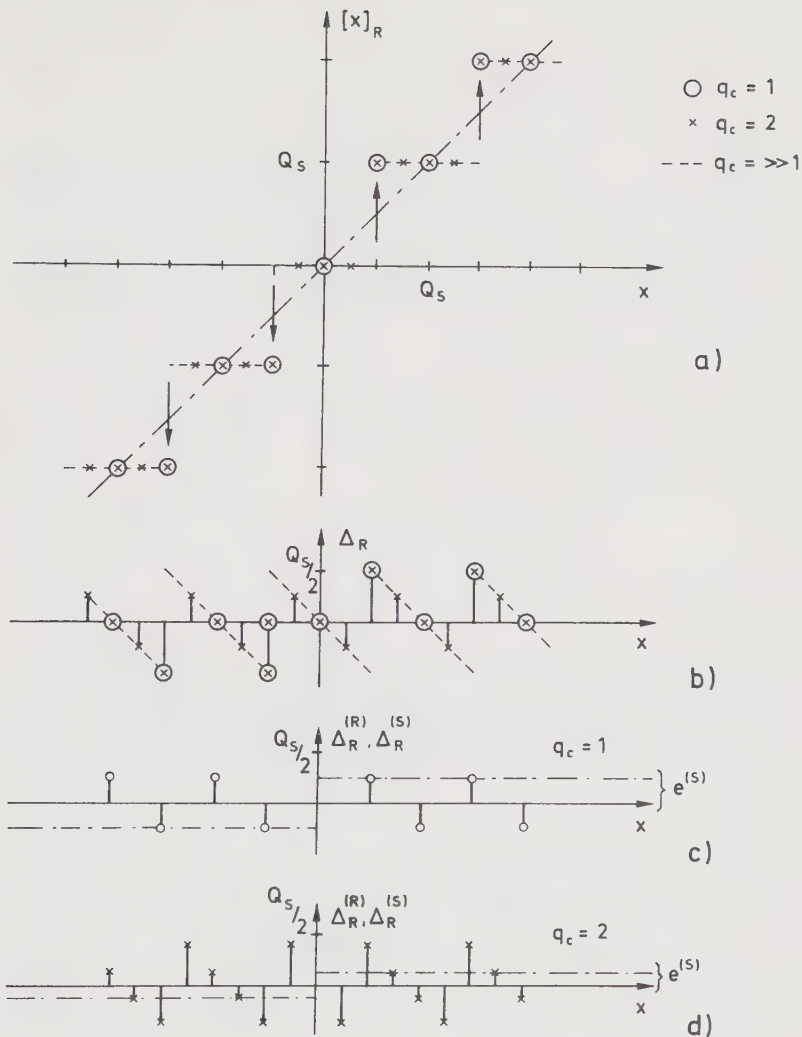


Bild 4.9: Quantisierungsdiagramm und Fehlerwerte bei betragsmäßigem Runden

Eine Variante des betragsmäßigen Rundens erhält man, wenn die Werte der halben Quantisierungsstufe nicht auf, sondern abgerundet werden. Es ergibt sich ein Unterschied im Vorzeichen des korrelierten Anteils.

4.2.2.3 Mathematisches Runden

Durch die Anwendung der strengen mathematischen Rundungsvorschrift ([70], Seite 97) kann auch bei einer Verkürzung um nur wenige Bit das Entstehen eines Gleichanteils oder eines mit dem Signal korrelierten Anteils vermieden werden. Die Werte der genau halben Quantisierungsstufe werden stets so gerundet, daß die nächstliegende Zahl der Quantisierung $2 Q_s$ erreicht wird (Bild 4.10a).

$$[x]_{MR} = \begin{cases} Q_s [Q_s^{-1} x + 0.5] & \text{für } x \neq x_v \\ 2 Q_s [\frac{Q_s^{-1}}{2} x + 0.5] & \text{für } x = x_v \end{cases} \quad (4.43)$$

mit $x_v = \pm (2v-1) \frac{Q_s}{2} \quad v = 1(1) Q_s^{-1}$

Im statistischen Mittel werden die Werte x_v nach Gl. (4.43) gleich häufig auf - bzw. abgerundet. Wie man anhand von Bild 4.10b sieht, verschwindet sowohl der lineare Mittelwert, als auch der mit dem Signal korrelierte Fehleranteil.

$$\overline{\Delta_{MR}} = e_{MR}^{(=)} = 0 \quad (4.44)$$

$$e_{MR}^{(S)} = 0 \quad (4.45)$$

Das mittlere Fehlerquadrat hat auch hier den Wert

$$\overline{\Delta_{MR}^2} = \sigma_{MR}^2 = \sigma_{MR}^{2(R)} = \frac{Q_s^2}{12} (1 + 2 Q_c^2) \quad (4.46)$$

und entspricht der Varianz des rauschartigen Fehlers.

4.2.2.4 Zweierkomplement-Abschneiden

Führt man die Verkürzung entsprechend Gl. (4.16) aus,

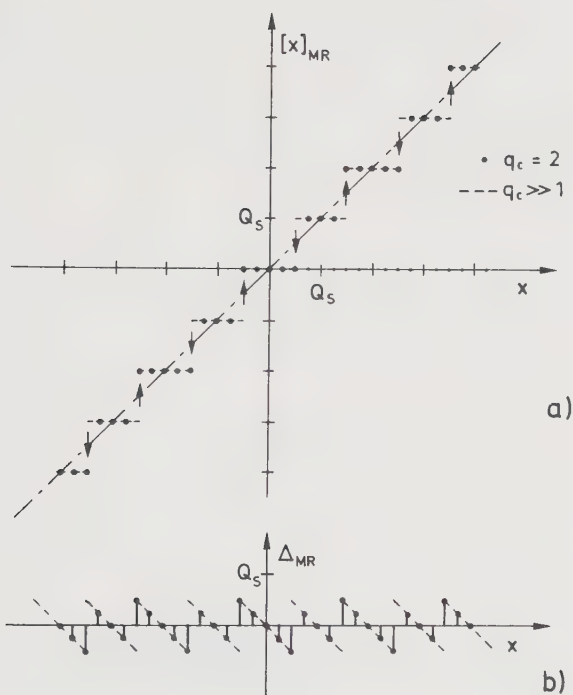


Bild 4.10: Quantisierungsdiagramm und Fehlerwerte bei mathematischem Runden

$$[x]_{TT} = Q_S [Q_S^{-1} x]$$

so ergeben sich bei Berücksichtigung des diskreten Modells entsprechend [7] die folgenden Kennwerte:

Gleichanteil

$$\overline{\Delta_{TT}} = e_{TT}^{(=)} = -\frac{1}{2} Q_S (1 - Q_C) , \quad (4.47)$$

mittleres Fehlerquadrat

$$\overline{\Delta_{TT}^2} = \frac{Q_S^2}{3} \left(1 - \frac{3}{2} Q_C + \frac{Q_C^2}{2} \right) \quad (4.48)$$

korrelierter Anteil

$$e_{TT}^{(S)} = 0 \quad (4.49)$$

Die Varianz des rauschartigen Fehleranteils berechnet sich somit zu

$$\sigma_{TT}^2 = \sigma_{TT}^{2(R)} = \overline{\Delta_{TT}^2} - \overline{\Delta_{TT}}^2 = \frac{Q_S^2}{12} (1 - Q_C^2) \quad (4.50)$$

Auch hier gehen die Kennwerte für $Q_C \rightarrow 0$ in die der entsprechenden kontinuierlichen Fehlerverteilung über.

4.2.2.5 Unsymmetrisches betragsmäßiges Abschneiden

Führt man die Quantisierung bei betragsmäßigem Abschneiden nach Abschnitt 4.2.1.3 in der Form durch, daß man zuerst ein Zweierkomplement-Abschneiden vornimmt und anschließend bei negativen Werten stets eine Quantisierungsstufe aufaddiert.

$$[x]_{UT} = Q_S ([Q_S^{-1} x] + \xi) \quad (4.51)$$

$$\begin{aligned} \text{mit} \quad \xi &= 0 & \text{für } x \geq 0 \\ \xi &= 1 & \text{für } x < 0 \end{aligned}$$

so werden bereits richtige Werte $x = v Q_S \quad v = 1(1) Q_S^{-1}$ weiter betragsmäßig verringert $|7|$. Diese zusätzlichen Fehler können bei einer kontinuierlichen Fehlerverteilung vernachlässigt werden. Bei einer diskreten Fehlerverteilung können sich hierdurch jedoch erhebliche Abweichungen ergeben. Für

$q_c = 1$ sind die möglichen Quantisierungen im Bild 4.11 eingetragen.

Aus der Lage der Fehlerwerte Δ_{UT} ($UT \hat{=}$ unsymmetrical truncation) im Bild 4.11b ersieht man, daß eine Aufteilung nach

$$\Delta_{UT} = \Delta_{UT}^{(=)} + \Delta_{UT}^{(S)} + \Delta_{UT}^{(R)} \quad (4.52)$$

möglich ist.

Es ergibt sich für die einzelnen Kennwerte:

Gleichanteil

$$\overline{\Delta_{UT}} = e_{UT}^{(=)} = \frac{Q_s Q_c}{2} , \quad (4.53)$$

mittleres Fehlerquadrat des Gesamtfehlers

$$\overline{\Delta_{UT}^2} = \frac{Q_s^2}{3} \left(1 + \frac{Q_c^2}{2} \right) , \quad (4.54)$$

Kennwert des korrelierten Anteils

$$e_{UT}^{(S)} = - \frac{Q_s}{2} \quad (4.55)$$

Die Varianz des rauschartigen Fehleranteils berechnet sich somit zu

$$\sigma_{UT}^{2(R)} = \overline{\Delta_{UT}^2} - e_{UT}^{(S)2} - \overline{\Delta_{UT}}^2 = \frac{Q_s^2}{12} (1 - Q_c^2) \quad (4.56)$$

4.2.2.6 Symmetrisches betragsmäßiges Abschneiden

Wird die im vorausgehenden Abschnitt beschriebene Korrektur für die Werte $x = -v Q_s$, $v = 1(1)Q_s^{-1}$ unterdrückt, so erhält man eine zum Nullpunkt symmetrische Quantisierungskennlinie.

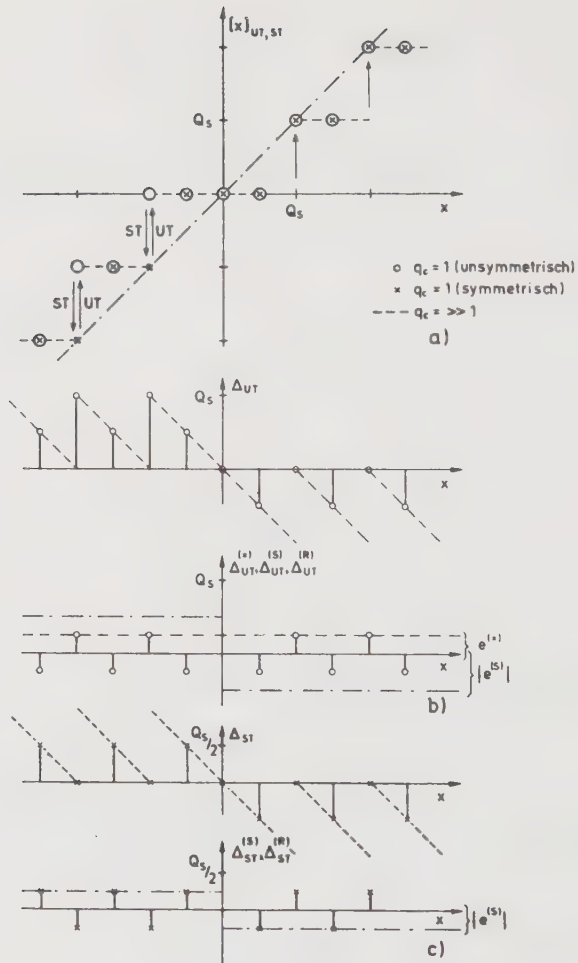


Bild 4.11: Quantisierungsdiagramm und Fehlerwerte bei unsymmetrischem und symmetrischem betragsmäßigem Abschneiden.

Für $q_c = 1$ sind die entsprechenden Werte mit × im Bild 4.11a gekennzeichnet.

Die dazugehörigen Fehlerwerte Δ_{ST} ($ST \hat{=}$ symmetrical truncation) sind dem Bild 4.11c zu entnehmen. Somit ergibt sich für die einzelnen Kennwerte:

Gleichanteil

$$\overline{\Delta_{ST}} = e_{ST}^{(=)} = 0 \quad (4.57)$$

mittleres Fehlerquadrat des Gesamtfehlers

$$\overline{\Delta_{ST}^2} = \frac{Q_S^2}{3} \left(1 - \frac{3}{2} Q_C + \frac{1}{2} Q_C^2 \right) \quad (4.58)$$

Kennwert des korrelierten Fehleranteils

$$e_{ST}^{(S)} = - \frac{Q_S}{2} (1 - Q_C) \quad (4.59)$$

Die Varianz des rauschartigen Anteils berechnet sich somit zu

$$\sigma_{ST}^{2(R)} = \overline{\Delta_{ST}^2} - (e_{ST}^{(S)})^2 = \frac{Q_S^2}{12} (1 - Q_C^2) \quad (4.60)$$

4.2.3 Berechnung der Rauschleistung eines Systems mit mehreren Nichtlinearitäten

Die Fehlerkennwerte der verschiedenen Nichtlinearitäten wurden in den vorausgehenden Abschnitten bestimmt. Mit Hilfe dieser Größen soll nun die Rauschleistung am Ausgang eines allgemeinen Übertragungssystems mit m Nichtlinearitäten (NL), wie es in Bild 4.12 dargestellt ist, bestimmt werden.

Nach Gl. (4.8) spaltet sich der Beitrag der einzelnen Nichtlinearitäten (Bild 4.13) in verschiedene Fehleranteile auf. Die rauschartigen, vom Signal unabhängigen Quellen $\sigma_i^{2(R)}$ sind dabei voneinander unabhängig. Während die mit dem Signal korre-

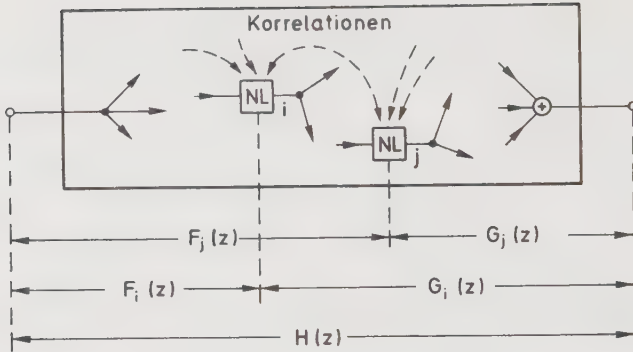


Bild 4.12: Allgemeines Übertragungssystem

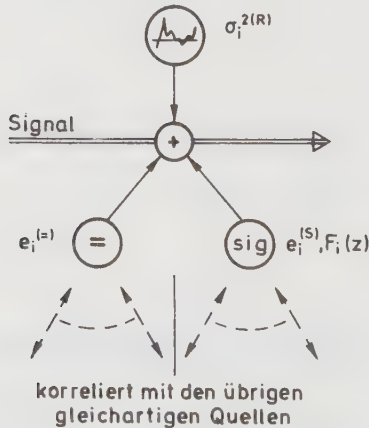


Bild 4.13: Modell einer allgemeinen Nichtlinearität

lierten Fehleranteile und die Gleichwerte verschiedener Nichtlinearitäten miteinander korreliert sind. Bei der Berechnung der Gesamtrauschleistung sind diese Korrelationen zu berücksichtigen [7], [58].

Der Übersicht halber sind die verschiedenen Kennwerte der einzelnen Quellen nochmals in Tabelle 4.1 zusammengestellt.

Art der Nicht-linearität	$e^{(=)}$	$e^{(S)}$	$\sigma^2(R)$	$\overline{\Delta^2}$
--------------------------	-----------	-----------	---------------	-----------------------

kontinuierliche Fehlerverteilung

Runden	R	O	O	$Q^2/12$	$Q^2/12$
Zweierkomplement-Abschneiden	TT	$-Q/2$	O	$Q^2/12$	$Q^2/3$
Betragsmäßiges Abschneiden	T	O	$-Q/2$	$Q^2/12$	$Q^2/3$

diskrete Fehlerverteilung

Zweierkomplement-Runden	TR	$\frac{Q_s Q_c}{2}$	O	$\frac{Q_s^2}{12} (1-Q_c^2)$	$\frac{Q_s^2}{12} (1+2Q_c^2)$
Betragsmäßiges Runden	R	O	$\frac{Q_s Q_c}{2}$	$\frac{Q_s^2}{12} (1-Q_c^2)$	$\frac{Q_s^2}{12} (1+2Q_c^2)$
Mathematisches Runden	MR	O	O	$\frac{Q_s^2}{12} (1+2Q_c^2)$	$\frac{Q_s^2}{12} (1+2Q_c^2)$
Zweierkomplement-Abschneiden	TT	$\frac{Q_s}{2} (1-Q_c)$	O	$\frac{Q_s^2}{12} (1-Q_c^2)$	$\frac{Q_s^2}{3} (1-\frac{3}{2}Q_c+\frac{Q_c^2}{2})$
Unsym.betrags. Abschneiden	UT	$\frac{Q_s Q_c}{2}$	$-\frac{Q_s}{2}$	$\frac{Q_s^2}{12} (1-Q_c^2)$	$\frac{Q_s^2}{3} (1+\frac{Q_c^2}{2})$
Sym. betrags. Abschneiden	ST	O	$-\frac{Q_s}{2} (1-Q_c)$	$\frac{Q_s^2}{12} (1-Q_c^2)$	$\frac{Q_s^2}{3} (1-\frac{3}{2}Q_c+\frac{Q_c^2}{2})$

Tabelle 4.1: Fehlerkennwerte der Nichtlinearitäten

Die Gesamtfehlerleistung N setzt sich nach Gl. (4.10)

$$N = N^{(R)} + N^{(S)} + N^{(=)} \quad (4.10)$$

aus dem rauschartigen $N^{(R)}$, dem mit dem Signal korrelierten $N^{(S)}$ und dem Gleichanteil $N^{(=)}$ zusammen.

4.2.3.1 Rauschartiger Fehleranteil

Das gleichverteilte Rauschsignal der i -ten Störquelle wird entsprechend der Übertragungsfunktion $G_i(z)$ zum Ausgang hin bewertet. Da die Anteile der verschiedenen Quellen voneinander unabhängig sind, können die Leistungen am Ausgang summiert werden.

$$N^{(R)} = \sum_{i=1}^m N_i^{(R)} = \frac{1}{2\pi j} \sum_{i=1}^m \sigma_i^2 \oint_{\Gamma} G_i(z) \cdot G_i(z^{-1}) \frac{dz}{z} \quad (4.61)$$

Die Berechnung des Ringintegrals kann mit Hilfe der Residuen-Rechnung durchgeführt werden.

4.2.3.2 Mit dem Signal korrelierter Fehleranteil

a) rauschartiges Eingangssignal

Unter der Voraussetzung eines rauschartigen gaußverteilten Eingangssignals der spektralen Leistungsdichte $\phi_{uu}(z)$ kann nach [58] die Leistungsdichte $\phi_{\Delta_y \Delta_y}^{(S)}$ des mit dem Signal korrelierten Fehleranteils am Ausgang des Filters berechnet werden. Da die verschiedenen Fehleranteile $\Delta_i^{(S)}$ auch untereinander korreliert sind, müssen neben den Leistungsspektren der einzelnen Fehlerquellen $\phi_{\Delta_{y_i} \Delta_{y_i}}^{(S)}(z)$ auch die Kreuzleistungsspektren der Fehlerquellen untereinander $\phi_{\Delta_{y_i} \Delta_{y_j}}^{(S)}(z)$ berücksichtigt werden.

Über die zweiseitige Z-Transformation erhält man die Auto- bzw. Kreuzkorrelationsfunktion des Signals am Eingang der Nichtlinearitäten

$$\varphi_{ij}(k) = \frac{1}{2\pi j} \oint_{\Gamma} F_i(z) \cdot F_j(z^{-1}) \phi_{uu}(z) z^{k-1} dz. \quad (4.62)$$

Da der korrelierte Fehleranteil der Signum-Funktion folgt, kann die Korrelationsfunktion des Fehlers $\varphi_{ij}^{(S)}$ am Ausgang der Nichtlinearität über die Polaritätskorrelationsfunktion

$$\varphi_{p_{ij}}(k) = \frac{2}{\pi} \arcsin \frac{\varphi_{ij}(k)}{\sqrt{\varphi_{ii}(0) \varphi_{jj}(0)}}$$

bestimmt werden. Mit den Kennwerten $e_i^{(S)}$ und $e_j^{(S)}$ erhält man

$$\begin{aligned} \varphi_{ij}^{(S)}(k) &= e_i^{(S)} e_j^{(S)} \varphi_{p_{ij}}(k) \\ &= e_i^{(S)} e_j^{(S)} \cdot \frac{2}{\pi} \arcsin \frac{\varphi_{ij}(k)}{\sqrt{\varphi_{ii}(0) \varphi_{jj}(0)}} \end{aligned} \quad (4.63)$$

Die spektrale Leistungs- bzw. Kreuzleistungsdichte $\phi_{\Delta_i \Delta_j}^{(S)}$ am Ausgang der Nichtlinearität berechnet sich anschließend durch Transformation

$$\phi_{\Delta_i \Delta_j}^{(S)} = Z \{ \varphi_{ij}^{(S)}(k) \} = \sum_{k=-\infty}^{+\infty} \varphi_{ij}^{(S)}(k) z^{-k}. \quad (4.64)$$

Die Leistungsdichten am Ausgang der Nichtlinearitäten werden dann noch mit den Übertragungsfunktionen bewertet und man erhält die Auto- bzw. Kreuzleistungsdichte des Fehlers am Ausgang des Filters zu

$$\phi_{\Delta_{Y_i} \Delta_{Y_j}}^{(S)}(z) = G_i(z) G_j(z) \phi_{\Delta_i \Delta_j}^{(S)} \quad (4.65)$$

Die korrelierte Fehlerleistung am Ausgang berechnet sich zu

$$N^{(S)} = \sum_{i=1}^m \sum_{j=1}^m N_{ij}^{(S)} = \frac{1}{2\pi j} \sum_{i=1}^m \sum_{j=1}^m \oint_{\Gamma} \phi_{\Delta_{Y_i} \Delta_{Y_j}}^{(S)}(z) \frac{dz}{z} \quad (4.66)$$

Die Auswertung der Gl. (4.30) ist in geschlossener Form nicht möglich. Die Berechnung kann jedoch numerisch mit Hilfe der diskreten Fourier-Transformation erfolgen. Der Rechenaufwand für zwei Abschneidstellen läßt sich dabei durch Verwendung von FFT-Algorithmen in noch vertretbaren Grenzen halten.

Er steigt jedoch quadratisch mit der Anzahl der Nichtlinearitäten. Für eine systematische Untersuchung einer größeren Anzahl von verschiedenen Filtersystemen, wie es z.B. bei der Optimierung der Anordnung der Teilsysteme einer Kaskade notwendig ist, eignet sich daher besser das vereinfachte linearisierte Modell, wie es in [7] und [58] vorgeschlagen wird. Der Rechenaufwand steigt dabei nur linear mit der Anzahl der Nichtlinearitäten und rechtfertigt aus dieser Sicht die durch die Näherung entstehenden Abweichungen. Eine Reihe berechneter Beispiele mit unterschiedlicher Anzahl von Nichtlinearitäten ergab im Mittel Abweichungen von 4 - 5 %.

Nach [58] ergibt sich für den mit dem Signal korrelierten Fehleranteil

$$N^{(S)} = \frac{1}{2\pi j} \oint_{\Gamma} \left(\frac{2}{\pi} \phi_{uu}(z) T(z) T(z^{-1}) + \left(1 - \frac{2}{\pi}\right) \sum_{i=1}^m (e_i^{(S)})^2 \cdot G_i(z) G_i(z^{-1}) \right) \frac{dz}{z} \quad (4.67)$$

mit

$$T(z) = \sum_{i=1}^m \frac{F_i(z) G_i(z)}{\sqrt{\psi_{ii}(0)}} (e_i^{(S)})^2$$

Die Autokorrelationsfunktion $\psi_{ii}(0) = f(\phi_{uu}, F_i)$ kann mit $i=j$ nach Gl. (4.62) bestimmt werden. Zur Reduzierung des Rechenaufwandes kann der vom Signal unabhängige Teil der Gl. (4.67) mit in die Berechnung des Rauschanteils $N^{(R)}$ einbezogen werden.

b) Sinusförmiges Eingangssignal

Im letzten Abschnitt wurde als Erregung des zu untersuchenden Systems ein gaußverteiltes Eingangssignal der spektralen Leistungsdichte $\phi_{uu}(\Omega)$ angenommen. Für das Beispiel einer betrags-

mäßigen Abschneidestelle wurde in Abschnitt 4.2.1.3 das Fehlerausgangsspektrum für verschiedene Eingangssignale gezeigt (Bild 4.6). Bei sinusförmigem Eingangssignal geht dabei der Spektralanteil des korrelierten Fehlers in ein diskretes Spektrum über. Die Fehlerleistung des mit dem Signal korrelierten Fehlers soll im folgenden für eine monofrequente Erregung berechnet werden.

Ausgehend von einer Sinusfolge der Frequenz f_1 erhält man bezogen auf die Abtastfrequenz f_A des diskreten Systems die normierte Signalfrequenz Ω_1

$$\Omega_1 = 2\pi \frac{f_1}{f_A} \quad (4.68)$$

Somit ergibt sich eine Eingangsfolge $\{u(k)\}$ mit

$$u(k) = \sin 2\pi k f_1 / f_A = \sin k \Omega_1 \quad (4.69)$$

Je nach Lage der Signalfrequenz f_1 zur Abtastfrequenz f_A entsteht bei der Abtastung eine periodische Folge unterschiedlicher Länge k_0 . Es gilt

$$u(k) \stackrel{!}{=} u(k+k_0)$$

und

$$\sin 2\pi k f_1 / f_A \stackrel{!}{=} \sin 2\pi (k+k_0) f_1 / f_A$$

Die Anzahl der Perioden des Sinussignals innerhalb der Signalperiode berechnet sich zu

$$k_1 = \frac{f_1}{f_A} \cdot k_0, \quad k_1 \in \mathbf{N} \quad (4.70a)$$

Für rationale Zahlen f_A/f_1 gilt somit auch

$$k_0 = k_1 \cdot \frac{f_A}{f_1}, \quad k_0 \in \mathbf{N} \quad (4.70b)$$

Für das Spektrum des Eingangssignals erhält man mit Gl. (4.68)

$$\Phi_{uu}(e^{j\Omega}) \stackrel{\hat{=}}{=} \Phi_{uu}(\Omega) = \begin{cases} 1, & \text{für } \Omega = \pm \Omega_1 \\ 0, & \text{sonst} \end{cases} \quad (4.71)$$

Auch bei sinusförmigem Eingangssignal ist eine Aufteilung des Fehlers in einen rauschartigen und einen mit dem Signal korrelierten Anteil möglich. Mit den in [55] für Runden durchgeführten Untersuchungen kann der rauschartige Anteil auch hier in einer Näherung als gleichverteilt angesehen werden. Diese Näherung gilt umso besser, je größer k_0 wird.

Für den mit dem Signal korrelierten Fehler $\Delta^{(S)}$ erhält man entsprechend Gl. (4.21) mit dem allgemeinen Kennwert $e^{(S)}$ sowie den Gln. (4.69) und (4.70a)

$$\Delta^{(S)}(k) = e^{(S)} \operatorname{sgn}(u(k)) = e^{(S)} \operatorname{sgn}(\sin 2\pi k \cdot \frac{k_1}{k_0}), \quad (4.72)$$

mit $\operatorname{sgn}(\sin 2\nu\pi) = 1$

und $\operatorname{sgn}(\sin(2\nu+1)\pi) = -1, \nu \in \mathbf{Z}$.

Das Fehlersignal ist ebenfalls periodisch. Es gilt

$$\Delta^{(S)}(k) = \Delta^{(S)}(k+k_0).$$

Bei der Berechnung des zugehörigen Fehlerspektrums kann man zunächst von dem entsprechend analogen Signal ausgehen und eine harmonische Analyse durchführen. Das so erhaltene Spektrum ist nicht bandbegrenzt. Bei der Abtastung des Signals entsteht somit eine Überfaltung des Spektrums. Verwendet man jedoch für die Berechnung der Spektralanteile die Beziehungen der diskreten Fourier-Analyse, so wird dieser Überfaltung automatisch Rechnung getragen [1]. Die Fourier-Koeffizienten c_ν berechnen sich hierbei zu

$$c_\nu = \frac{1}{k_0} \sum_{k=0}^{k_0-1} \Delta^{(S)}(k) e^{-j \frac{2\pi \nu k}{k_0}}, \quad \nu = 0(1)k_0-1$$

und mit Gl. (4.72)

$$c_\nu = \frac{e^{(S)}}{k_0} \sum_{k=0}^{k_0-1} \operatorname{sgn}(\sin 2\pi k \frac{k_1}{k_0}) \cdot e^{-j \frac{2\pi \nu k}{k_0}}, \quad (4.73)$$

$$\nu = 0(1)k_0-1.$$

Das Leistungsspektrum des korrelierten Fehlers kann mit Hilfe der Fourier-Koeffizienten berechnet werden.

$$\phi_{\Delta\Delta}^{(S)}(\Omega) = \begin{cases} \frac{|c_v|^2}{2}, & \text{für } \Omega = \Omega_1 \frac{v}{k_1} \quad v = 0(1)k_0-1 \\ 0, & \text{sonst} \end{cases} \quad (4.74)$$

Durch Summation der einzelnen Spektralanteile erhält man die Leistung des korrelierten Fehlers,

$$\overline{\Delta^{(S)2}} = \frac{1}{2} \sum_{v=0}^{k_0-1} |c_v|^2. \quad (4.75)$$

Die Anzahl der Fourier-Koeffizienten steigt mit k_0 . Für große Werte k_0 , bzw. $k_0 \rightarrow \infty$ ist eine Berechnung der Spektralanteile durch Faltung der Fourier-Reihe der Signum-Funktion mit einem Impulskamm des Abstands $\Omega_A = 2\pi$ möglich. Die Fourier-Reihe kann je nach gewünschter Genauigkeit nach einer endlichen Anzahl von Werten abgebrochen werden.

Enthält ein System mehrere mit dem Signal korrelierte Fehlerquellen, so sind auch hier die Korrelationen der verschiedenen Nichtlinearitäten zu berücksichtigen.

Die Fehlerfolgen der einzelnen Nichtlinearitäten werden entsprechend der Phasenlagen ψ_i der Signale an diesen Stellen erzeugt und mit der Übertragungsfunktion $G_i(z)$ zum Ausgang hin übertragen.

$$\psi_i = -\angle F_i(z = e^{j\Omega_1}) \quad , \quad i = 1(1)m \quad (4.76a)$$

Die komplexen Amplituden des Fehlers der i -ten Nichtlinearität am Ausgang des Systems erhält man mit Gl. (4.73)

$$c_{y_i v} = c_{i v} \cdot e^{-j\psi_i} \cdot G_i(z = e^{j \frac{2\pi v k}{k_0}}) \quad (4.76b)$$

Die Gesamtleistung des korrelierten Fehleranteils bei m Nichtlinearitäten errechnet sich mit Gln. (4.76) zu

$$N^{(S)} = \frac{1}{2} k_o \sum_{v=0}^{m-1} \left| \sum_{i=1}^m c_{y_{iv}} \right|^2 \quad (4.77)$$

Mit Gl. (4.77) ist es nun möglich, auch bei sinusförmigem Eingangssignal eine Fehleranalyse für betragsmäßiges Abschneiden durchzuführen. Die Abhängigkeit der Rauschleistung in Abhängigkeit von der Eingangsfrequenz wird an einem Beispiel im Abschnitt 6.3 gezeigt und mit den Simulationsergebnissen verglichen.

4.2.3.3 Gleichanteil des Fehlers

Der Gleichanteil der einzelnen Nichtlinearitäten ist vom Signal unabhängig. Zur Bestimmung des gesamten Gleichanteils am Ausgang werden die einzelnen Werte mit den Übertragungsfunktionen $G_i(z=1)$ bewertet und aufsummiert. Die Leistung des Gleichanteils $N^{(=)}$ berechnet sich zu

$$N^{(=)} = \left(\sum_{i=1}^m e_i^{(=)} G_i(z=1) \right)^2 \quad (4.78)$$

In den vorausgehenden Abschnitten wurden die verschiedenen Fehlerleistungsanteile eines Systems mit mehreren Quellen bestimmt. Somit liegt ein umfassendes analytisches Modell zur Bewertung der verschiedenen digitalen Schaltungen vor.

4.3 Dynamikbereich eines digitalen Systems

In den vorausgehenden Abschnitten wurden analytische Modelle zur Berechnung des Quantisierungsgeräusches angegeben. Für die Bewertung einer Struktur ist aber nicht die absolute Rauschleistung, sondern das Rausch-Signal-Verhältnis von Interesse. Eine Optimierung des Dynamikbereiches der einzelnen Strukturen ist daher notwendig.

Bei der Festkomma-Darstellung der Zustandsvariablen ist der Wertevorrat betragsmäßig nach oben hin beschränkt (Abschnitt 4.1). Eine Überschreitung des Zahlenbereichs führt zu nicht-linearen Verzerrungen, die mit Sicherheit zu großen Fehlern, gegebenenfalls auch zu Überlaufgrenzyklen führen können [1]. Um einen Überlauf im System zu vermeiden, werden daher bei vorgegebenem Eingangssignal die verschiedenen Summationsstellen innerhalb des Filters und dessen Ausgang überprüft und entsprechende Skalierungsfaktoren eingeführt [3], [4], [6], [71]-[74].

4.3.1 Notwendige Prüfstellen

Bei einer Darstellung der Zahlen im Zweierkomplement, die hier stets angenommen wird, kann die Überprüfung des Dynamikbereichs auf die Summationspunkte unmittelbar vor den Multiplizierern und auf den Ausgang des Systems beschränkt werden [75]. Eine Übersteuerung der Teilsummen ist erlaubt. Zur Veranschaulichung lassen sich nach [4] die Zahlenwerte in Zweierkomplement auf einem geschlossenen Ring anordnen. Eine Übersteuerung der Teilsummen, d.h. ein "Überrunden des Zahlenbereichs" kann durch eine entsprechende "Summation in Gegenrichtung" wieder ausgeglichen werden.

Diese, aus der Literatur bekannte, mögliche Übersteuerung der Teilsummen läßt sich auf erlaubte Übersteuerungen der einzelnen Multiplizierer erweitern.

Geht man davon aus, daß die Koeffizienten mit $|a| > 1$ in der Form

$$a = a' \cdot 2^p, \quad -1 \leq a' < 1 \quad (4.79a)$$

dargestellt werden, so kann bei dem Teilprodukt

$$z' = a' \cdot x, \quad -1 \leq x < 1 \quad (4.79b)$$

selbst keine Übersteuerung auftreten. Die anschließende Multiplikation mit einer Zweierpotenz (Schiebeoperation) stellt eine 2^p -fache Addition des Teilprodukts z' zur Gesamtsumme dar. Eine Übersteuerung dieser Teilsummen ist zulässig, wenn bei der Summation mehrerer Produkte nur die Gesamtsumme betragsmäßig < 1 ist.

4.3.2 Aussteuerungskriterien

Eine obere Schranke für die maximale Amplitude läßt sich zur Vermeidung jeglicher Übersteuerung an den vorausgehend beschriebenen kritischen Stellen abschätzen. Unter der Voraussetzung, daß die Rundungsfehler vernachlässigbar klein gegenüber dem Signal sind, erhält man am Ausgang einer Übertragungsstrecke mit der Impulsantwort $h(k)$

$$y(k) = \sum_{\kappa=0}^{\infty} h(k) \cdot u(k-\kappa) \quad (4.80)$$

Ist die Eingangsfolge $\{u(k)\}$ nach Gl. (4.3) auf Werte $|u(k)| \leq 1$ beschränkt, so ergibt sich für die maximale Amplitude des Ausgangssignals $|1|$

$$|y(k)| \leq \sum_{k=0}^{\infty} |h(k)| \quad (4.81)$$

Aufgrund der oberen Schranke nach Gl.(4.81) kann man einen Skalierungsfaktor K finden, damit gilt

$$\sum_{k=0}^{\infty} |h'(k)| = \sum_{k=0}^{\infty} |K h(k)| \leq 1 \quad (4.82)$$

und die Ausgangsfolge $\{y(k)\}$ wiederum auf Werte $|y(k)| \leq 1$ beschränkt bleibt.

Der Skalierungsfaktor errechnet sich zu

$$K = \frac{1}{\sum_{k=0}^{\infty} |h(k)|} \quad (4.83)$$

Wird die Skalierung entsprechend dem absoluten Kriterium nach Gl. (4.82) durchgeführt, so stellt sich bei der Simulation des Systems heraus, daß diese Grenze in den meisten Fällen zu pessimistisch gewählt ist und somit das Rausch-Signalverhältnis unnötig verschlechtert.

Für die Beurteilung unterschiedlicher Strukturen geeignete Aussteuerungsgrenzen werden von Jackson in [71] und [72] angegeben. Durch Z-Transformation der Gl. (4.80) erhält man unter der Voraussetzung absolut summierbarer Eingangssignale $\{u(k)\}$,

$$Y(z) = H(z) U(z) \quad (4.84)$$

und mit $z = e^{j\Omega}$ durch Rücktransformation

$$y(k) = \frac{1}{2\pi} \int_0^{2\pi} H(\Omega) U(\Omega) e^{jk\Omega} d\Omega \quad (4.85)$$

Abgeleitet aus der Holder'schen Ungleichung ergibt sich mit Hilfe der L_p -Norm ¹⁾ eine allgemeine Abschätzung des maximalen Betrags der Ausgangsfolge $\{y(k)\}$.

$$|y(k)| \leq \|H\|_p \cdot \|U\|_q \quad (4.86)$$

mit $\frac{1}{p} + \frac{1}{q} = 1$ und $p, q \geq 1$.

Unter der Voraussetzung, daß das Eingangsspektrum $U(\Omega)$ die Norm $\|U\|_q \leq 1$, für $q \geq 1$ erfüllt und weiter für die Ausgangsfolge $\{y(k)\}$ $|y(k)| \leq 1$ gelten soll, ist die Skalierung entsprechend

$$\|H'\|_p = \|K H\|_p = K \|H\|_p \leq 1 \quad (4.87)$$

mit $p = q/(q-1)$ durchzuführen.

¹⁾ L_p -Norm: $\|A\|_p = \frac{1}{2\pi} \left[\int_0^{2\pi} |A(\Omega)|^p d\Omega \right]^{1/p}$

mit dem Grenzwert

$$\|A\|_\infty = \max_{0 \leq \Omega \leq 2\pi} |A(\Omega)|$$

Der Skalierungsfaktor errechnet sich für unterschiedliche Normen zu

$$K = \|H\|_p^{-1} \quad (4.88)$$

Für die praktische Anwendung bei der Realisierung digitaler Systeme eignen sich besonders die Normen (p,q) gleich $(1,\infty)$, $(2,2)$ und $(\infty,1)$ [72]. Mit $p = \infty$ wird dabei die schärfste Bedingung an die Übertragungsfunktion gestellt. Man erhält mit

$$\|y(k)\|_1 \leq \|H\|_\infty \cdot \|u\|_1 = K \|H\|_\infty \cdot \|u\|_1 \leq 1 \quad (4.89)$$

eine Bedingung, die auch für sinusförmige Eingangssignale der Form $u(k) = \hat{U} \sin(k\Omega_1)$ gilt [72]. Dabei ist

$$\|u\|_1 = \hat{U} \quad (4.90)$$

Mit Gl. (4.89) und (4.90) erhält man für $\hat{U} \leq 1$ den Aussteuerungsfaktor

$$K = \|H\|_\infty^{-1} \quad (4.91)$$

Überträgt man die Beziehungen (4.86) auf die Autokorrelationsfunktion bzw. die Varianz eines rauschartigen Eingangssignals mit dem Spektrum $\phi_{uu}(z)$, so kann mittels der L_2 -Norm eine Aussage über die mittlere Leistung an den Übersteuerungspunkten gemacht werden. Nach [72] erhält man

$$\sigma_y^2 \leq \|H\|_p^2 \cdot \|\phi_{uu}\|_q = \|H\|_{2p}^2 \cdot \|\phi_{uu}\|_q, \quad (4.92)$$

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Mit den Parametern $p = 1$ und $q = \infty$ ergibt sich

$$\sigma_y^2 \leq \|H\|_2^2 \cdot \|\phi_{uu}\|_\infty = K^2 \|H\|_2^2 \cdot \|\phi_{uu}\|_\infty \quad (4.93)$$

Für ein "gaußverteiltes" Eingangssignal der Varianz σ_u^2 kann eine Schranke $n\sigma_u$ gewählt werden, die in Abhängigkeit von n mit einer bestimmten Wahrscheinlichkeit nicht überschritten wird. Setzt man nun in Gl. (4.93):

$$\|\phi_{uu}\|_{\infty} \stackrel{!}{=} n^2 \sigma_u^2$$

und bestimmt K in der Art, daß $\sigma_y^2 = \sigma_u^2$ gilt, so wird die Schranke $n\sigma_y$ am Ausgang des Systems im zeitlichen Mittel nicht häufiger überschritten als vom Eingangssignal selbst. Der Skalierungsfaktor ergibt sich somit zu

$$K = \|\mathbf{H}\|_2^{-1} \quad (4.94)$$

Ersetzt man die Übertragungsfunktion $H(z)$ durch die entsprechende Übertragungsfunktion $F_1(z)$ vom Eingang zu den Übersteuerungspunkten, so können mit den Gln. (4.83), (4.91) und (4.94) die notwendigen Skalierungsfaktoren einer Struktur bestimmt werden.

Der Einfluß der Wahl der Skalierung auf das Rausch-Nutzleistungsverhältnis wird am Beispiel eines reinen Nennerblocks 2. Grades mit einer Rundungsstelle am Ausgang im Bild 4.14 gezeigt.

Für den gewählten Fall gilt:

$$H(z) = F(z) = G(z) = \frac{1}{z^2 + c_1 z + c_0}$$

mit $c_0 = \rho_P^2 = 0.81$ und $c_1 = -2\rho_P \cos \psi_P$.

Speist man die Schaltung mit weißem Rauschen der Varianz $\sigma_u^2=1$, so erhält man für die Nutzleistung P am Ausgang des Filters

$$P = K^2 \oint_{\Gamma} H(z) \cdot H(z^{-1}) \frac{dz}{z} .$$

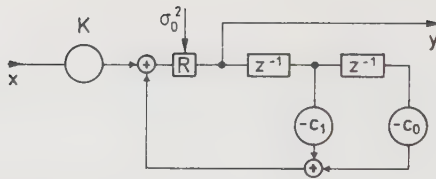
Die Rundungsstelle verursacht einen rauschartigen Fehler der Varianz $\sigma_o^2 = Q^2/12$ und die Fehlerleistung N am Ausgang beträgt

$$N = \sigma_o^2 \oint_{\Gamma} G(z) G(z^{-1}) \frac{dz}{z} .$$

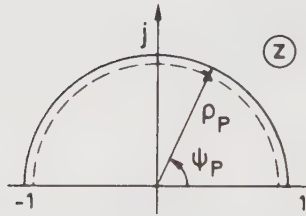
Mit $G(z) = H(z)$ erhält man somit das Rauschnutzleistungsverhältnis zu

$$N/P = \sigma_o^2 / K^2$$

Das Rausch-Nutzleistungsverhältnis ist also hier umgekehrt proportional zum Quadrat von K, das nach Gln.(4.83), (4.91) und (4.94) gewählt werden kann. Die Ergebnisse sind im Bild 4.14c dargestellt.



a)



b)

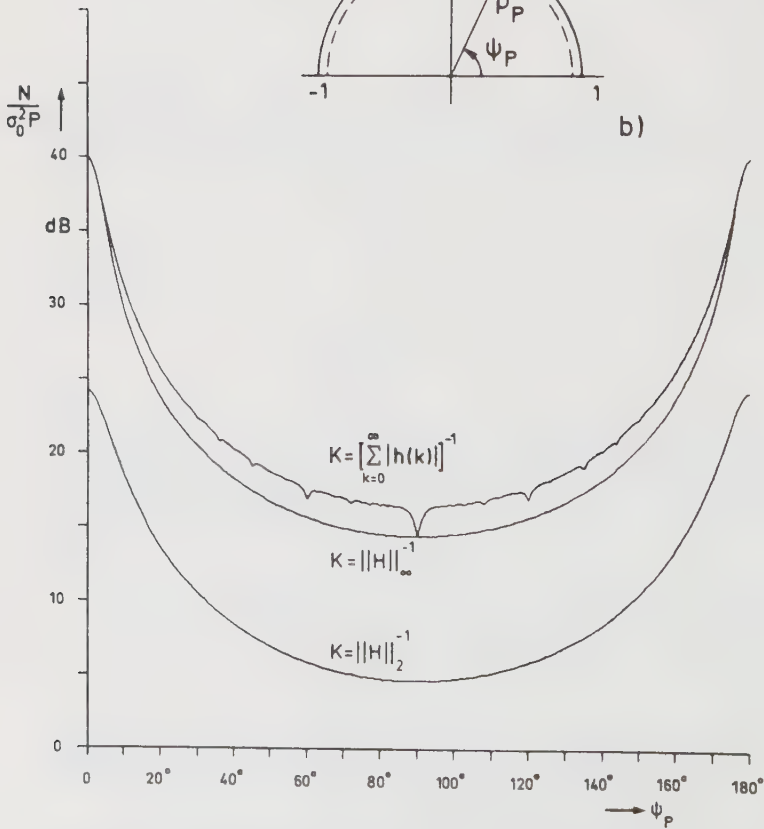


Bild 4.14: Rausch-Nutzleistung bei verschiedener Skalierung in Abhängigkeit vom Polwinkel ψ_P
a) Schaltung, b) Wahl der Pöllage,
c) Rauschverhalten.

4.4 Rauschzahl und zusätzliche innere Wortlänge Δw

Für die Berechnung der Rauschleistung N am Ausgang eines digitalen Systems wurden im Abschnitt 4.2 analytische Modelle angegeben. Die Rauschleistung war dabei stets eine Funktion der Teilübertragungsfunktionen und der Quantisierungsstufe Q bzw. Q_s der Zustandsvariablen. Durch eine Verfeinerung der Stufe Q_s kann daher das Quantisierungsrauschen in allen digitalen Systemen beliebig reduziert werden. Zum Vergleich der verschiedenen Strukturen wurde daher in [1] die Rauschleistung am Ausgang eines Systems auf die Rauschleistung einer einzelnen Rundungsstelle bei kontinuierlichen Signalen $Q_s^2/12$ bezogen. Man erhält die Rauschzahl R ,

$$R = \frac{12}{Q_s^2} N = 12 \cdot 2^{2q_s} N \quad (4.95)$$

Werden in einem System unterschiedliche Quantisierungen der Zustandsvariablen verwendet, so können für die einzelnen nicht miteinander korrelierten Quantisierungsstellen Teilrauschzahlen R_j definiert werden. Dabei können Nichtlinearitäten gleicher Quantisierung zusammengefaßt werden. Für die Gesamtleistung gilt dann,

$$N = \frac{1}{12} \sum_j Q_{sj}^2 R_j \quad (4.96)$$

Der Rauschleistungsanteil der größten Quantisierung bildet für große Unterschiede Δq_s eine untere Rauschleistungsgrenze. Es erscheint daher sinnvoll, die Gesamtrauschzahl auf die größte Quantisierungsstufe zu beziehen.

$$R = \frac{1}{\max_j \{Q_{sj}\}^2} \sum_j Q_{sj}^2 R_j \quad (4.97)$$

Da durch die größte Quantisierungsstufe im System, die in der Regel im vorgeschalteten A/D-Wandler liegt, eine untere Schranke für das Rauschen vorgegeben ist, erscheint es nicht zweckmäßig, die Wortlänge im Innern eines Teilsystems beliebig zu erhöhen. Als Beispiel wird ein Filtersystem nach Bild 4.15 gewählt. Die allgemeinen Parameter Q_{sj} und R_j werden dabei entsprechend den

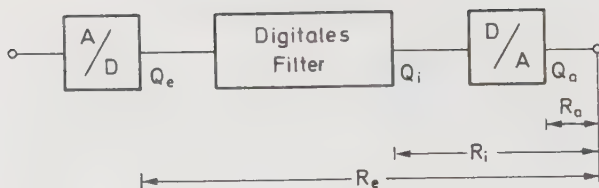


Bild 4.15: Filter mit A/D- und D/A-Wandler

Funktionsblöcken bezeichnet. Der A/D-Wandler hat eine Quantisierungsstufe Q_e und sein Rauschen wird mit R_e bewertet. Das Filter hat eine Quantisierungsstufe Q_i und eine Rauschzahl R_i , der abschließende D/A-Wandler Q_a und R_a . Liegen die Werte für R_e und R_a sowie für Q_e und Q_a fest, so wird R_i und Q_i in $|1|$ entsprechend der Bedingung

$$R_i Q_i^2 \leq R_e Q_e^2 + R_a Q_a^2 \quad (4.98)$$

gewählt. Liegt nun auch R_i für eine vorgegebene Schaltung fest, so kann eine maximale Quantisierung Q_i mit Gl. (4.98) bzw. eine minimal notwendige Länge $w_i = q_i + 1$ der Zustandsvariablen im Filter angegeben werden. Der Unterschied zur kleinsten fest vorgegebenen Wortlänge $w_{\min} = q_{\min} + 1$ wird mit zusätzliche notwendige Wortlänge Δw bezeichnet. Mit der größten fest vorgegebenen Quantisierungsstufe

$$Q_{\max} = \max \{Q_e, Q_a\}$$

erhält man $q_{\min} = -\lg Q_{\max}$. Die notwendige innere Wortlänge, die stets ganzzahlig sein muß, ergibt sich mit Gl. (4.98) zu

$$w_i = 1 + \left\langle \frac{1}{2} \lg \frac{R_i}{R_e Q_e^2 + R_a Q_a^2} \right\rangle 1, \quad (4.99)$$

1) $\langle a \rangle$ derart, daß $0 \leq \langle a \rangle - a < 1$

die zusätzliche Wortlänge Δw zu

$$\Delta w = w_i - w_{\min} = \left\langle \frac{1}{2} \lg \frac{R_i Q_{\max}^2}{R_e Q_e^2 + R_a Q_a^2} \right\rangle \quad (4.100)$$

Für den in der Regel vorliegenden Fall mit $Q_a = Q_e = Q_{\max}$ und $R_a = 1$ vereinfacht sich der Ausdruck für Δw . Es gilt dann

$$\Delta w = \left\langle \frac{1}{2} \lg \frac{R_i}{R_e + 1} \right\rangle . \quad (4.101)$$

Die zusätzliche Wortlänge Δw stellt ein geeignetes Maß für die Bewertung des von der Zustandsvariablenlänge abhängenden Realisierungsaufwandes dar. Zur Minimierung des Aufwandes ist also unter sonst gleichwertigen Strukturen die zu wählen, bei der R_i und somit Δw minimal werden.

Ein allgemeines Aufwandsmaß A zur Beurteilung unterschiedlicher Realisierungen wurde in Abschnitt 3.3.1 eingeführt und mit Gl. (3.17) in Teilfaktoren aufgespalten. Die Bewertung des Aufwandes A_S in Abhängigkeit von den Zustandsvariablen wird hier entsprechend der Bewertung des Koeffizientenaufwandes auf einen theoretischen Mindestaufwand bei minimaler Wortlänge $w_{\min}$ bezogen. Unter der stark vereinfachenden Annahme, daß der Realisierungsaufwand proportional zur Länge der Zustandsvariablen ist, erhält man mit Gl. (4.100) für den relativen Aufwandsfaktor A_S

$$A_S = \frac{w_i}{w_{\min}} = 1 + \frac{\Delta w}{w_{\min}} \quad (4.102)$$

Somit kann die Bewertung des Rauschverhaltens mit in das allgemeine Aufwandsmaß nach Gl. (3.17) aufgenommen werden und im Zusammenhang mit Koeffizientenaufwand und der Filtergraderweiterung betrachtet werden.

5. Analytischer Vergleich des Rauschverhaltens unterschiedlicher Realisierungen

Neben den seit längerer Zeit bekannten Strukturen der ersten und zweiten kanonischen Form [1], den gekoppelten Realisierungen nach [2] und den Grundstrukturen nach Jackson [71] sowie deren Zusammenschaltung zu Kaskaden- oder Parallelanordnungen wurden in letzter Zeit weitere neue Strukturen vorgeschlagen [6], [41], [62], [64], [76] - [86]. Berücksichtigt man darüber hinaus die Möglichkeit, alle diese Strukturen wiederum mit unterschiedlichen Arithmetiken realisieren zu können, so ergibt sich eine auf den ersten Blick unübersehbare Menge unterschiedlicher Strukturen. Die Aufgabe der folgenden Abschnitte soll es nun sein, durch Auswahl eines Teils der möglichen Strukturen und deren Vergleich einen Überblick über das Rauschverhalten zu geben.

5.1 Übersicht über die untersuchten Strukturen

5.1.1 Einführung

Die Untersuchung der Koeffizientenempfindlichkeit nach [6], [23], [37], [41] - [47], [87] hat zumindest für reguläre ¹⁾ Strukturen ergeben, daß es sinnvoll erscheint, höhergradige Systeme in Teilblöcken zweiten Grades zu realisieren. Die günstigen Eigenschaften der direkten Realisierung bei nichtrekursiven Filtern [53] in Bezug auf das Rundungsrauschen werden bei den entsprechenden rekursiven Systemen durch die ungünstigen Aussteuerungsbedingungen aufgehoben. Darüber hinaus erscheint eine Gliederung in Teilblöcke auch für den Schaltungsaufbau günstig.

¹⁾ regulär: Ein Netzwerk vom Grad n wird in [85] regulär genannt, wenn es Rückführungsschleifen vom Grad 1 bis n enthält.

Die Untersuchung des Rauschverhaltens unterschiedlicher Teilblöcke zweiten Grades ergibt einen groben Überblick über deren Verwendbarkeit in höhergradigen Systemen und kann so zu einer ersten Selektion bei der Auswahl einer geeigneten Struktur für die Realisierung einer vorgegebenen Übertragungsfunktion dienen. Bei der Vielzahl der möglichen Strukturen erscheint eine vollständige Behandlung aller aus der Literatur bekannten Strukturen unmöglich. Anhand einiger ausgewählter Beispiele sollen jedoch die prinzipiellen Unterschiede im Rauschverhalten gezeigt werden. Die Signalflußgraphen dieser Strukturen werden im nachfolgenden Abschnitt zusammengestellt. Das Rauschverhalten dieser Grundstrukturen bei unterschiedlicher Arithmetik wird dann im Abschnitt 5.2, das der Zusammenschaltung zu höhergradigen Filtern im Abschnitt 5.3 näher untersucht.

Zur eindeutigen Kennzeichnung werden im folgenden die verschiedenen Grundstrukturen mit arabischen Ziffern bezeichnet, die Realisierungen mit unterschiedlichen Arithmetiken durch Buchstaben verdeutlicht.

5.1.2 Katalog der untersuchten Strukturen

a) Kanonische Strukturen nach Hwang [85]

Von Hwang werden in [85] 14 reguläre und in Bezug auf Zustandspeicher, Multiplizierer und Addierer kanonische Schaltungen angegeben. Die beiden ersten Schaltungen stimmen dabei mit den kanonischen nach [1] überein. Die Überprüfung der Schaltungen im Hinblick auf ihre Realisierbarkeit ergibt, daß die 8., 9. und 14. Form (nach Hwang) wegen verzögerungsfreier Schleifen nicht realisierbar sind. Für einen allgemeinen Block 2. Grades scheinen nur die ersten vier Strukturen geeignet zu sein (Tab. 5.1)

	Lit.	Untersuchungen	Bemerkungen
STRUKTUR 1	[1] — [4] [6] [71] [85]	Runden, (Abschn. 5.2.3) Bild: 5.10, 5.12, 5.14 Runden vor und nach der Zusammenfassung der Teilprodukte (Abschn. 5.2.4.1) Bild: 5.30, 5.31)	kanonisch in Bezug auf Multiplizierer, Addierer und Zustandsspeicher voneinander unabhängige Zähler- und Nennerkoeffizienten; Einsparung einer Multiplikation bei Nullstelle auf dem Einheitskreis
STRUKTUR 2 $H(z) = \frac{b_2 z^2 + b_1 z + b_0}{z^2 + c_1 z + c_0}$	[86]	Zweierkomplement Abschneiden (Abschn. 5.2.4.2) Bild 5.33, 5.34) Betragsmäßiges Abschneiden (Abschn. 5.2.4.3), Bild 5.38, 5.39 Skalierung mit 2er-Potenzen (Abschn. 5.2.7) Bild 5.51	Struktur 1 $\hat{=}$ 2D Struktur 2 $\hat{=}$ 1D nach Jackson [71]
STRUKTUR 3	[86] [87]	Runden nach möglicher Zusammenfassung der Teilprodukte (Abschn. 5.2.3) Pol-Nullstellen-Test, Bild 5.10	3. und 4. kanonische Struktur nach Hwang [86] kanonisch in Bezug auf Multiplizierer Addierer und Zustandsspeicher Zählerkoeffizient. von Nennerkoeff. abhängig
STRUKTUR 4 $H(z) = \frac{h_2 z^2 + (h_2 c_1 + h_1) z + (h_2 c_0 + h_1 c_1 + h_0)}{z^2 + c_1 z + c_0}$		Allpaßtest Bild 5.12 Abtastintervall-Test, Bild 5.14	keine Einsparung am Multiplizierer keine Zwischenspeicher notwendig

Tabelle 5.1: Grundstrukturen I (nach Hwang)

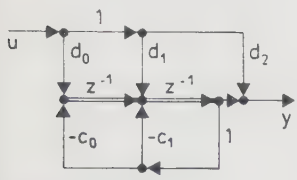
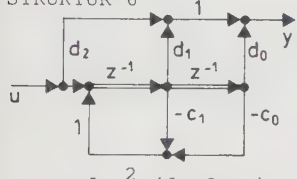
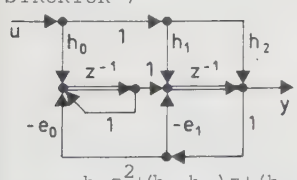
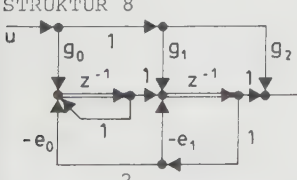
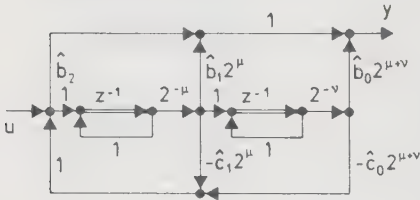
	Lit.	Untersuchungen	Bemerkungen
STRUKTUR 5 	71 77	Runden nach mögl. Zusammenfassungen (Abschn. 5.2.3) Pol-Nullstellen-Test Bild 5.11	Strukturen nach Jackson 71 kanonisch in Bezug auf Multiplizierer, Addierer und Zustandsspeicher für $d_2=b_2=1$ (keine Skalierung) zeigen die Strukturen die gleiche Koeffizientenempfindlichkeit wie die Strukturen 1 und 2
STRUKTUR 6  $H(z) = \frac{d_2 z^2 + (d_1 + d_2 c_1) z + (d_0 + d_2 c_0)}{z^2 + c_1 z + c_0}$		Allpaßtest Bild 5.13 Abtastintervall-Test Bild 5.15	Zählerkoeff. von Nennerkoeff. abhängig. keine Einsparung an Multiplizierern möglich
STRUKTUR 7  $H(z) = \frac{h_2 z^2 + (h_1 - h_2) z + (h_0 - h_1)}{z^2 + (e_1 - 1) z + (e_0 - e_1)}$	77	Runden nach mögl. Zusammenfassungen der Teilprodukte (Abschn. 5.2.3) Pol-Nullstellen-Test Bild 5.11 Allpaß-Test Bild 5.13	Strukturen nach Parker u. Hess kanonisch in Bezug auf Multiplizierer und Zustandsspeicher sonst wie oben
STRUKTUR 8  $H(z) = \frac{g_2 z^2 + (g_1 + g_2 e_1 - g_2) z + (g_0 - g_1 + g_2 e_0 - g_2 e_1)}{z^2 + (e_1 - 1) z + (e_0 - e_1)}$		Abtastintervall-Test Bild 5.15	

Tabelle 5.2: Grundstrukturen II

Struktur

Bemerkungen

STRUKTUR 9



$$H(z) = \frac{\tilde{b}_2 z^2 + (\tilde{b}_1 - 2\tilde{b}_2)z + (\tilde{b}_0 - \tilde{b}_1 + \tilde{b}_2)}{z^2 + (\tilde{c}_1 - 2)z + (\tilde{c}_0 - \tilde{c}_1 + 1)}$$

Struktur nach [41]

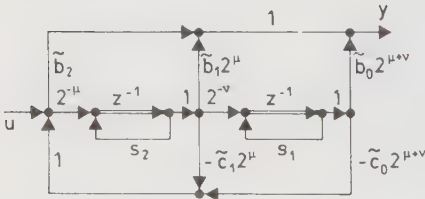
kanonisch in Bezug auf Multiplizierer

Zustandsspeicher verlängert
folglich höhere Realisierungskosten
zusätzliche "Shift"-Operationen

Untersuchung des Rauschverhaltens
siehe Abschnitt: 5.2.6

Bild 5.47, 5.48

STRUKTUR 10



$$H(z) = \frac{\tilde{b}_2 z^2 + (\tilde{b}_1 - (s_1 + s_2)\tilde{b}_2)z + (\tilde{b}_0 - s_1\tilde{b}_1 + s_1s_2\tilde{b}_2)}{z^2 + (\tilde{c}_1 - s_1 - s_2)z + (\tilde{c}_0 - s_1\tilde{c}_1 + s_1s_2)}$$

In Bezug auf die Zustandsspeicher kanonische Struktur mit

$$s_1, s_2 \in \{-1, 0, +1\}$$

kanonisch in Bezug auf Multiplizierer

zusätzliche "Shift"-Operationen

Untersuchung des Rauschverhaltens
siehe Abschnitt: 5.2.6

Bild 5.47, 5.48

Tabelle 5.3: Grundstrukturen III

(mit rückgekoppelten Speichern und inneren Skalieren)

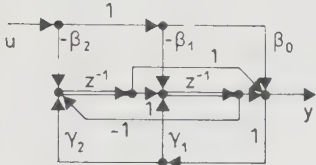
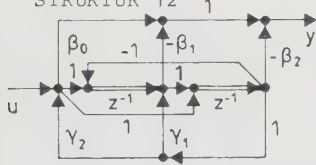
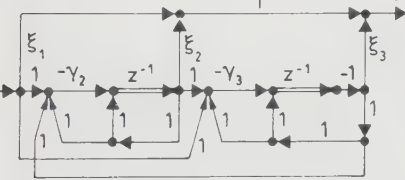
S t r u k t u r	Lit.	B e m e r k u n g e n
STRUKTUR 11  STRUKTUR 12 	[62] [62]	Struktur nach Meerkötter und Wegener kanonisch in Bezug auf Multiplizierer und Zustandsspeicher Zähler- und Nennerkoeffizienten voneinander unabhängig mit betragsmäßigem Abschneiden in der Nennerschleife Grenzzyklen bei verschwindendem Eingangssignal Untersuchung des Rauschverhaltens siehe Abschnitt 5.2.5 Bild 5.43
$H(z) = \frac{\beta_0 z^2 - (\beta_1 + \beta_2) z + (\beta_0 + \beta_1 - \beta_2)}{z^2 - (\gamma_1 + \gamma_2) z + (1 + \gamma_1 - \gamma_2)}$		
STRUKTUR 13 	[64]	Struktur nach Verkroost und Butterweck kanonisch in Bezug auf Multiplizierer und Zustandsspeicher Zähler- und Nennerkoeffizienten voneinander abhängig mit "kontrolliertem Runden" in jeder Hinsicht grenzzyklenfrei Untersuchung des Rauschverhaltens siehe Abschnitt 5.2.5 Bild 5.43 Nicht für alle Pol-Nullstellenlagen realisierbar
$H(z) = \frac{\xi_1 z^2 + b_1 z + b_0}{z^2 + (\gamma_2 - \gamma_3) z + (\gamma_2 + \gamma_3 - 1)}$ $b_1 = 2((\gamma_2 + \gamma_3 - 1) \xi_1 - \gamma_2 \xi_2 - \gamma_3 \xi_3)$ $b_0 = 2(\xi_1 (\gamma_2 - \gamma_3) - \gamma_2 \xi_2 + \gamma_3 \xi_3)$		

Tabelle 5.4: Grundstrukturen IV (Grenzzyklen freie Strukturen)

	Lit.	Untersuchungen	Bemerkungen
STRUKTUR 14 $H(z) = \frac{(g_2 + g_1 + g_0)z^2 + (g_1 + g_0)c_1z + g_0c_0}{z^2 + c_1z + c_0}$	[86] [87]	Pol-Nullstellentest bei Runden nach mögl. Zusammenfassung der Teilprodukte (Abschn. 5.2.3) Bild 5.22 5.24	entspricht der 6. kanon. Struktur nach Hwang [86] nicht für alle Pol-lagen realisierbar $c_1 \neq 0, c_0 \neq 0$
STRUKTUR 15 $H(z) = \frac{b_2z^2 + b_1z + b_0}{z^2 + c_1z + c_0}$ mit $b_2 = h_0; \quad b_1 = \frac{h_0}{h_1h_2} + \frac{h_0}{h_2h_3} + \frac{h_0}{h_3h_4} + \frac{1}{h_1}; \quad b_0 = \frac{h_0}{h_1h_2h_3h_4} + \frac{1}{h_1h_3h_4} + \frac{1}{h_1h_2h_3}$ $c_1 = \frac{1}{h_1h_2} + \frac{1}{h_2h_3} + \frac{1}{h_3h_4}; \quad c_0 = \frac{1}{h_1h_2h_3h_4}$	[86] [87] [78]	Diagramme von Linien gleicher Rauschzahlen Abschn. 5.2.3 Bild 5.22 5.25	entspricht der 11. kanonischen Struktur nach Hwang [86] nicht für alle Pol-Nullstellenlagen realisierbar

Tabelle 5.5: Grundstrukturen V (nach Hwang)

Struktur	Lit.	Untersuchungen	Bemerkungen
16 - "coupled biquad" $H(z) = \frac{a_2 z^2 + (a_1 - a_2) g_1 z + (a_2 - a_0) g_1 h_1}{z^2 + g_1 z + g_1 h_1}$	[80]	Runden nach mögl. Zusammenfassung der Teilprodukte Abschn. 5.2.3 Bild 5.23	kanonisch in Bezug auf Multiplizierer Addierer und Zustandsspeicher nicht für alle Pol-lagen realisierbar $c_1, c_0 \neq 0$
17 - Brückenstruktur nach Markel und Gray $H(z) = \frac{b_2 z^2 + b_1 z + b_0}{z^2 + b_0 (1 + k_1) z + k_1}$	[81] [82]	Runden nach mögl. Zusammenfassung der Teilprodukte Abschn. 5.2.3 Bild 5.23	kanonisch in Bezug auf Multiplizierer und Zustandsspeicher für $k_0, k_1 < 1$ stets stabil nicht für alle Nullstellenlagen realisierbar ungünstig für "Pipeline"-Betrieb.

$$b_2 = a_0 + (a_0 + a_2)k_1 + (a_0 + a_1)k_0(1+k_1) ; b_1 = (1+k_1)(a_1 + a_2k_0) ; b_0 = a_2$$

Tabelle 5.6: Grundstrukturen VI (Brückenstrukturen)

Die 5., 6., 11., 12. und 13. Form nach Hwang sind nur bedingt realisierbar, da für bestimmte Pol-Nullstellenlagen einzelne Koeffizienten über alle Grenzen anwachsen. Als Beispiele für das ungünstige Rauschverhalten dieser Strukturen werden die 6. und 11. näher untersucht. Ihre Signalflußgraphen finden sich in Tabelle 5.5.

b) Strukturen mit gleicher Koeffizientenempfindlichkeit

Von Jackson in [71] und Parker, Hess in [77] wurde eine Reihe von Strukturen angegeben, die für $b_2 = 1$ alle das gleiche Koeffizientenverhalten aufweisen. Ein Teil dieser Strukturen wurde bereits von Jackson [71] als für allgemeine Anwendungen weniger geeignet ausgemustert. Dies wurde durch eigene Untersuchungen bestätigt. Ebenso kann ein großer Teil der Strukturen nach [77] bei Minimierung der Rauschquellen für die weiteren Betrachtungen ausgeschlossen werden. Die neben der 1. und 2. kanonischen Form am geeignetsten erscheinenden Strukturen sind in Tabelle 5.2 zusammengestellt.

c) Strukturen mit rückgekoppelten Speichern und inneren Skalierern

Von Agarwal und Burrus wurden in [41] Strukturen angegeben, die für Pol- und Nullstellen in der Gegend $z = 1$ ein günstiges Rauschverhalten zeigen. Die dort angegebenen Schaltungen sind kanonisch in Bezug auf die Multiplizierer und die Anzahl an Zustandsspeichern. Zur Vermeidung von Überläufen wird die Wortlänge der Zustandsspeicher im Vergleich zu den Zustandsvariablen der Multiplizierer verlängert. Die in [41] mit 2A bezeichnete Struktur ist in Tabelle 5.3, Struktur 9 dargestellt.

Durch eine Verschiebung der Skalierungsmultiplikationen und der anschließenden Nichtlinearitäten vor die Zustandsspeicher wird eine neue Lösung mit gleicher Zustandsvariablenwortlänge für Speicher und Multiplizierer gefunden (Struktur 10). Bei einer positiven Rückkopplung der Speicher (Integrierer) ist eine Rauschverbesserung auf das Gebiet um $z = 1$ beschränkt. Kehrt man das Vorzeichen bei der Rückführung um, so erhält man eine entsprechende Verbesserung der Rauscheigenschaften für Pole- und Nullstellen in der Gegend von $z = -1$.

Allgemein kann man durch eine freie Wahl der Rückkopplungen zusammen mit einfachen Speichern im ganzen neun unterschiedliche Schaltungen erstellen, die durch einfaches Umschalten realisiert werden können (Bild 5.0).

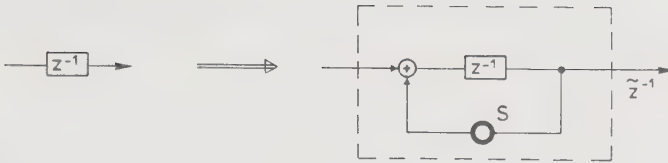


Bild 5.0: Transformation der Zustandsspeicher

$s = +1$ Integrierer $s = 0$ einfacher Speicher
 $s = -1$

$$\tilde{z} = z - s \quad (5.1)$$

Eine Optimierung der Struktur durch Wahl der jeweils günstigsten Parameter s_1, s_2 ist möglich.

d) Grenzzyklenfreie Struktur

Von Meerkötter und Wegener wurden in [62] Strukturen angegeben, die bei unerregtem Eingang aufgrund der "Pseudo-Passivität" der Schaltung [68] gänzlich grenzzyklenfrei sind. Die Strukturen 11 und 12 in Tabelle 5.4 sind zueinander transponiert.

Eine weitere Verbesserung des Grenzzyklenverhaltens erhält man mit dem von Verkroost und Butterweck in [64] vorgeschlagenen Block 2. Grades mit gesteuerten Nichtlinearitäten. Mit der in Tabelle 5.4, Struktur 13, gezeigten Schaltung gelingt es, bei entsprechender Rundung auch bei konstantem Eingangssignal Grenzzyklen zu vermeiden.

e) Kanonische Leitungs- und Brückenstrukturen

Von verschiedenen Autoren werden eine Reihe von Leitungs- und Brückenstrukturen vorgeschlagen. Die Verwendbarkeit für den Aufbau von allgemeinen Blöcken 2. Grades wurde in den Vorarbeiten überprüft. Als Beispiele werden in der Tabelle 5.5 die

sogenannte "coupled biquad"-Form nach [80] und eine der möglichen "one multiplier" Brückenstrukturen nach Markel und Gray [81], [82] angegeben.

5.2 Untersuchung der Grundblöcke 2. Grades

Bei der Untersuchung der Teilblöcke wird von der Übertragungsfunktion

$$H(z) = \frac{b_2 z^2 + b_1 z + b_0}{z^2 + c_1 z + c_0} = b_2 \frac{z^2 + \frac{b_1}{b_2} z + \frac{b_0}{b_2}}{z^2 + c_1 z + c_0} \quad (5.2)$$

ausgegangen. Die Wahl des Faktors b_2 soll dabei stets so erfolgen, daß die Gesamtübertragungsfunktion entsprechend der L_∞ -Norm (siehe Abschnitt 4.3) ausgesteuert wird. Es gilt also $\|H\|_\infty = 1$. Zur optimalen Aussteuerung der Teilschaltungen ist entsprechend Abschnitt 4.3.1 eine Skalierung der Struktur vorzunehmen. Nach Bild 5.2 wird der Gesamtfaktor b_2 aufgeteilt

$$b_2 = b_3 \cdot b_2' \cdot b_4 \quad (5.3a)$$

$$b_1' = b_1 \cdot b_2' / b_2 ; \quad b_0' = b_0 \cdot b_2' / b_2 \quad (5.3b)$$

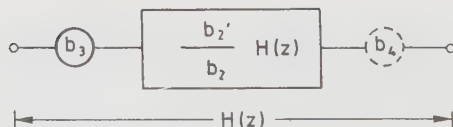


Bild 5.1: Skalierungsfaktoren

Zur Wahl der Faktoren b_3 , b_2' und b_4 können die Grundstrukturen in zwei Gruppen aufgeteilt werden. Bei der ersten Gruppe (Struktur 1, 3, 5, 7, 8, 11) hängen alle Koeffizienten der Teilübertragungsfunktionen $F_i(z)$ zu den Skalierungspunkten von einem gemeinsamen Zählerverstärkungsfaktor b_2 bzw. b_2' ab und die Aussteuerung kann allein durch diese Faktoren erfolgen. Mit K_i nach Gl. (4.91) erhält man für die m Skalierungs-

punkte

$$b_2' = b_2 \cdot \min \{K_i\}, \quad i = 1(1)m \quad (5.4)$$

$$b_3 = 1$$

Im Gegensatz hierzu steht die zweite Gruppe (Strukturen 2, 4, 6, 9, 10, 12, 13), bei der nicht alle Übertragungsfunktionen von einem gemeinsamen Faktor abhängen. Die Einführung eines Vorfaktors b_3 wird notwendig. Werden mit m_u die Anzahl der unabhängigen und mit m_a die Zahl der von b_2' abhängigen Aussteuerungsfaktoren bezeichnet, so ergibt sich

$$b_3 = \min \{K_i\} \quad i = 1(1)m_u \quad (5.5a)$$

$$b_2' = \frac{b_2}{b_3} \cdot \min \{K_j\} \quad j = 1(1)m_a \quad (5.5b)$$

Für $b_3 \cdot b_2' < b_2$ wird die Struktur am Ausgang nicht voll angesteuert. Da die notwendige Anhebung des Signals stets in einem nachfolgenden Schaltungsteil erfolgen kann, wird hierfür kein zusätzlicher Multiplizierer vorgesehen. Zur Bestimmung des Rauschsignalleistungsverhältnisses wird jedoch der Faktor b_4

$$b_4 = b_2 / (b_2' b_3) \quad (5.6)$$

benötigt. Da bei vorgegebenem Eingangssignal ein übersteuerungsfreier Betrieb gegebenenfalls nur mit dem Vorfaktor b_3 möglich ist, wird dieser, soweit notwendig, mit in die Untersuchungen einbezogen.

Für die Berechnung der Rauschleistung und Rauschzahlen der einzelnen Strukturen wird davon ausgegangen, daß alle Verkürzungen der Zustandsvariablen auf eine einheitliche Wortlänge $w_s = q_s + 1$, der Wortlänge der internen Zustandsspeicher vorgenommen werden.

5.2.1 Beschreibung der durchgeführten Tests

Zum Vergleich der Strukturen werden die Rauschzahlen nach Gl. (4.95) als Funktion der Pol- und Nullstellen bestimmt. Da eine gleichzeitige Überprüfung beider Abhängigkeiten schwer darstellbar ist, werden im folgenden drei Tests ausgewählt.

a) Pol-Nullstellen-Test

Die Lage der Nullstellen des Übertragungssystems wird hierbei fest auf dem Einheitskreis $\rho_N=1$ vorgegeben. Die Wahl des Nullstellenwinkels erfolgte willkürlich zu $\psi_N = \pm 30^\circ$. Das Rauschverhalten der Struktur wird dann in Abhängigkeit vom Polwinkel ψ_P bei konstantem Polradius ρ_P entsprechend Bild 5.2 untersucht. Die Übertragungsfunktion ergibt sich in Abhängigkeit von den Parametern zu

$$H(z) = b_2 \cdot \frac{z^2 - 2\cos 30^\circ z + 1}{z^2 - 2\rho_P \cos \psi_P z + \rho_P^2} \quad (5.7)$$

Die Rauschzahlen werden über dem Polwinkel aufgetragen (siehe z.B. Bild 5.10).

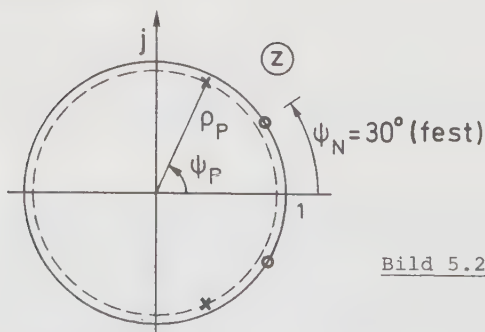


Bild 5.2: Pol-Nullstellen-Test

b) Allpaß-Test

Zur Bestimmung des Allpaßverhaltens werden die Pole des Blockes 2. Grades ebenfalls auf einem konstant vorgegebenen Radius verschoben. Die zugehörigen Nullstellen werden entsprechend der Allpaßbedingung $|1|$ gewählt. Man erhält für die Übertragungs-

funktion

$$H(z) = \frac{c_0 z^2 + c_1 z + 1}{z^2 + c_1 z + c_0} = \frac{\rho_P^2 z^2 - 2\rho_P \cos \psi_P z + 1}{z^2 - 2\rho_P \cos \psi_P z + \rho_P^2} \quad (5.8)$$

Da es speziell für die Realisierung von Allpässen Strukturen gibt, die mit nur zwei Multiplizierern aufgebaut werden können (z.B. [88]), ist der Allpaß-Test direkt nur für programmierbare Filter von Interesse.

c) Abtastintervall-Test

Die Anwendung digitaler Filter zur Simulation analoger Systeme [89] stellt die Frage nach einer geeigneten Struktur für diese Aufgabe. Die Pol-Nullstellenlage des Blockes 2. Grades wird im analogen Bereich so gewählt, daß bei bilinearer Transformation mit dem Abtastintervall $T/T_0 = 1$ im digitalen Bereich ein Polpaar bei $\rho_P = 0,9$, $\psi_P = \pm 20^\circ$ und ein Nullstellenpaar bei $\rho_N = 1$, $\psi_N = \pm 30^\circ$ entsteht. Zur Überprüfung des Rauschverhaltens werden nun die digitalen Blöcke 2. Grades bei Transformationen mit unterschiedlichen Abtastintervallen im Bereich $0 \leq T/T_0 \leq 1$ untersucht. Die Rauschzahlen werden im logarithmischen Maßstab über T/T_0 aufgetragen.

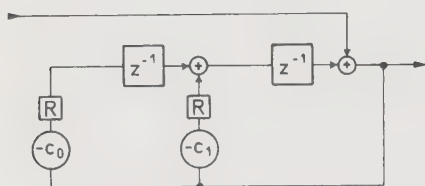
5.2.2 Rauschverhalten rein rekursiver Strukturen

Untersucht wird das Rauschverhalten des rein rekursiven Teils der Grundstruktur in der ersten und zweiten kanonischen Form. Hierdurch wird der prinzipielle Einfluß der Nennerübertragungsfunktion auf die Rauschquellen (R) verdeutlicht. Die Aussteuerung der Struktur bleibt hier unbeachtet. Mit Gl. (4.61) und (4.95) und den Fehlerübertragungsfunktionen

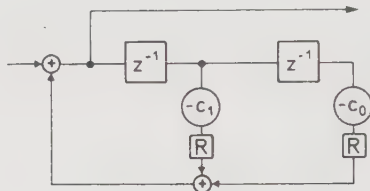
$$G(z) = H(z) = \frac{1}{z^2 + c_1 z + c_0} = \frac{1}{z^2 - 2\rho_P \cos \psi_P z + \rho_P^2}$$

erhält man für die Rauschzahlen bei 2 Rauschquellen

$$R(\rho_P, \psi_P) = \frac{2(1+c_0)}{(1-c_0)[(1+c_0)^2 - c_1^2]} = \frac{2(1+\rho_P^2)}{(1-\rho_P^2)[(1+\rho_P^2) - 4\rho_P^2 \cos^2 \psi_P]} \quad (5.9)$$



1. kan. Form



2. kan. Form

Bild 5.3: Nenner-Blöcke

Entsprechend dem Pol-Nullstellentest werden in Bild 5.4a die Rauschzahlen in Abhängigkeit vom Polwinkel ψ_P bei konstantem Radius ρ_P aufgetragen. Um den gleichzeitigen Einfluß von Radius und Winkel beobachten zu können, werden in Bild 5.4b Linien gleicher Rauschzahlen aufgetragen. Aus den Darstellungen wird das starke Ansteigen der Rauschleistung in der Nähe des Einheitskreises ersichtlich. Für $\psi_P \rightarrow 0, 180^\circ$ wirkt sich darüber hinaus die Annäherung der beiden Pole verstärkend aus. Die dreidimensionale Abbildung der Rauschzahlen über der Koeffizientenebene $c_1 = -2\rho_P \cos \psi_P$, $c_0 = \rho_P^2$ macht diese Abhängigkeiten weiter deutlich (Bild 5.4c).

5.2.3 Das Rauschverhalten unterschiedlicher Strukturen bei Runden

Die verschiedenen, im Abschnitt 5.1.2 vorgestellten Grundstrukturen werden im folgenden miteinander verglichen. Bei der Festlegung der Rundungsstellen werden die unverkürzten Teilprodukte soweit möglich (und sinnvoll) vor der Rundung zusammengefaßt.

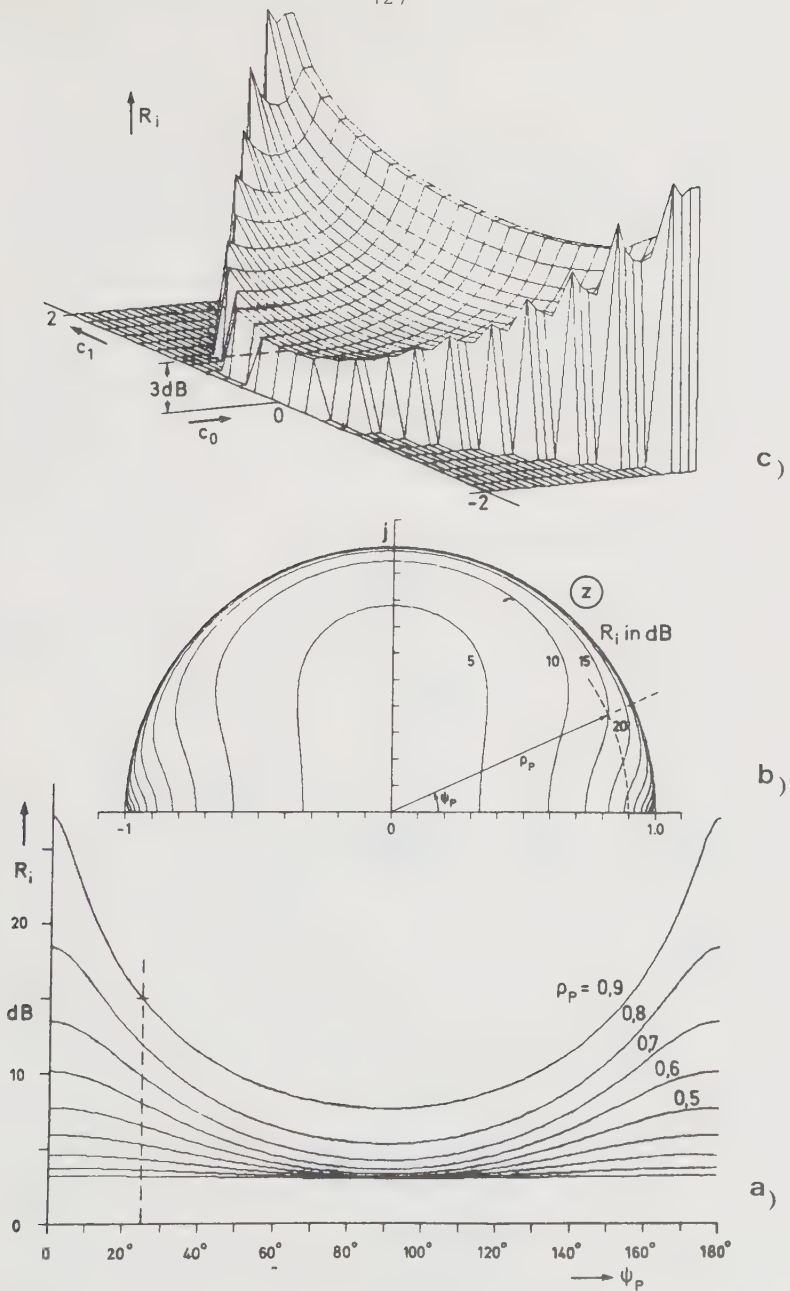


Bild 5.4: Verschiedene Darstellungen des Rauschverhaltens reiner Nenner-Blöcke 2. Grades

So können z.B. bei der 2. kanonischen Form (Tab. 5.1) die sonst vorhandenen 6 Rauschquellen auf 2 reduziert werden. Bei einer seriellen Verarbeitung der Zustandsvariablen wird für diese Maßnahme kein nennenswerter schaltungstechnischer Mehraufwand benötigt (|39|, |1|). Die Strukturen werden mit "b" gekennzeichnet.

Die Skalierung der Schaltungen wird entsprechend Abschnitt 4.3.2 nach der L_∞ -Norm mit Gl. (4.91) vorgenommen. Die Berechnung der Rauschleistung erfolgt nach Abschnitt 4.2.3.1, Gl. (4.61).

5.2.3.1 Katalog der Ergebnisse

a) Struktur 1 - 1. kanonische Form

Signalflußgraph: Tabelle 5.1

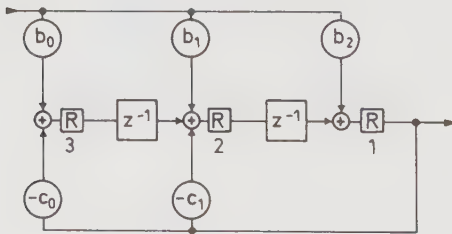


Bild 5.5: Blockschaltbild Struktur 1 b

Übertragungsfunktion:

$$H(z) = \frac{b_2 z^2 + b_1 z + b_0}{z^2 + c_1 z + c_0} = \frac{B(z)}{C(z)} \quad (5.10)$$

Skalierung: ${}_1F_1(z) = H(z) \quad (5.11)$

$$b_3 = 1 ; \quad b'_2 = b_2 ; \quad b_4 = 1 \quad (5.12)$$

Fehlerübertragungsfunktionen:

$${}_1G_1(z) = z^2/C(z)$$

$${}_1G_2(z) = z/C(z) \quad (5.13)$$

$${}_1G_3(z) = 1/C(z)$$

Pol-Nullstellen-Test: Bild 5.10

Allpaß-Test : Bild 5.12

Abtastintervall-Test: Bild 5.14

b) Struktur_2_-2_kanonische_Form

Signalflußgraph: Tabelle 5.1

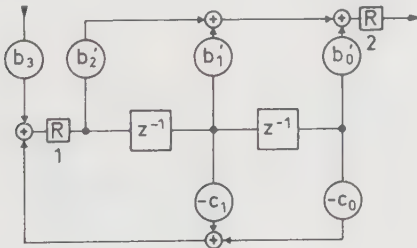


Bild 5.6: Blockschaltbild Struktur 2 b

Übertragungsfunktion: ${}_2H(z) = {}_1H(z)$ nach Gl. (5.10)

Skalierung: ${}_2F_1(z) = z^2/C(z) \quad (5.14)$

$$b_3 = ||{}_2F_1||_{\infty}^{-1} ; \quad (5.15)$$

$$b_2' = b_2/b_3 ; \quad b_4 = 1$$

$$b_i' = b_i/b_3 ; \quad i = 0, 1 \quad (5.16)$$

Fehlerübertragungsfunktionen:

$${}_2G_1(z) = \frac{1}{b_3} H(z) ; \quad {}_2G_2(z) = 1 \quad (5.17)$$

Pol-Nullstellen-Test: Bild 5.10

Allpaß-Test : Bild 5.12

Abtastintervall-Test: Bild 5.14

c) Struktur 3 - 3. kanonische Form nach Hwang

Signalflußgraph: Tabelle 5.1

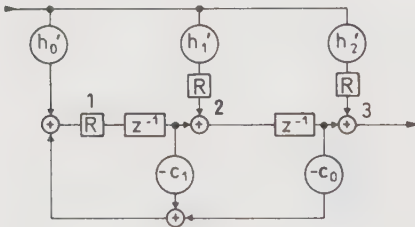


Bild 5.7: Blockschaltbild Struktur 3b

Übertragungsfunktion:
$${}_3H(z) = \frac{h_2 z^2 (h_2 c_1 + h_1) z + (h_2 c_0 + h_1 c_1 + h_0)}{z^2 + c_1 z + c_0} \quad (5.18)$$

Koeffizienten unskaliert :

$$\begin{aligned} h_2 &= b_2 \\ h_1 &= b_1 - h_2 c_1 \\ h_0 &= b_0 - h_2 c_0 - h_1 c_1 \end{aligned} \quad (5.19)$$

Skalierung:

$$\begin{aligned} {}_3F_1(z) &= \frac{h_0 z^2 - c_0 h_1 z}{C(z)} \\ {}_3F_2(z) &= \frac{h_1 z^2 + (h_0 + c_1 h_1) z}{C(z)} \\ {}_3F_3(z) &= {}_3H(z) \end{aligned} \quad (5.20)$$

$$b'_2 = b_2 \left[\max_i ||{}_3F_i||_\infty \right]^{-1}, \quad i = 1, 2, 3$$

$$b_3 = 1 ; \quad b_4 = b_2/b_2'$$

$$h_i' = h_i \cdot b_2'/b_2, \quad i = 0, 1, 2 \quad (5.21)$$

Fehlerübertragungsfunktionen:

$${}_3G_1(z) = b_4/C(z) ; \quad {}_3G_2(z) = b_4z/C(z)$$

$${}_3G_3(z) = b_4 \quad (5.22)$$

Pol-Nullstellen-Test: Bild 5.10

Allpaß-Test : Bild 5.12

Abtastintervall-Test: Bild 5.14

d) Struktur 4 - 4. kanonische Form nach Hwang

Signalflußgraph: Tabelle 5.1

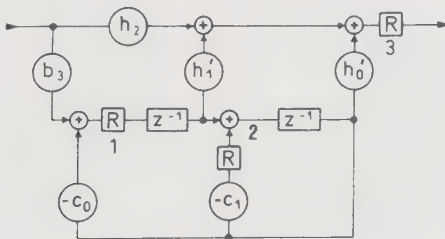


Bild 5.8: Blockschaltbild Struktur 4b¹⁾

Übertragungsfunktion: ${}_4H(z) = {}_3H(z)$ nach (5.18)

Koeffizienten unskaliert: siehe Gl. (5.19)

Skalierung: ${}_4F_1(z) = \frac{z^2 + c_1z}{C(z)} ; \quad {}_4F_2(z) = z/C(z)$

$${}_4F_3(z) = {}_4H(z) \quad (5.23)$$

¹⁾ Vorfaktor b_3 wurde so gewählt, daß eine minimale Anzahl von Rundungsstellen erreicht wird.

$$b_3 = \left[\max_i \{ \| {}_4F_i \|_\infty \} \right]^{-1} \quad i = 1, 2$$

$$b_2' = b_2/b_3 ; \quad b_4 = 1$$

$$h_i' = h_i/b_3 \quad i = 0, 1 \quad (5.24)$$

Fehlerübertragungsfunktionen:

$${}_4G_1(z) = (h_1'z + (h_0' + h_1'c_1))/C(z)$$

$${}_4G_2(z) = (h_0'z - h_1'c_0)/C(z)$$

$${}_4G_3(z) = 1 \quad (5.25)$$

Pol-Nullstellen-Test: Bild 5.10

Allpaß-Test : Bild 5.12

Abtastintervall-Test: Bild 5.14

e) Struktur 5 - 2C nach Jackson

Signalflußgraph: Tabelle 5.2

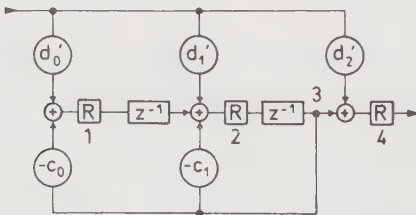


Bild 5.9: Blockschaltbild Struktur 5 b

Übertragungsfunktion:

$${}_5H(z) = \frac{d_2z^2 + (d_1+d_2c_1)z + (d_0+d_2c_0)}{z^2 + c_1z + c_0} \quad (5.26)$$

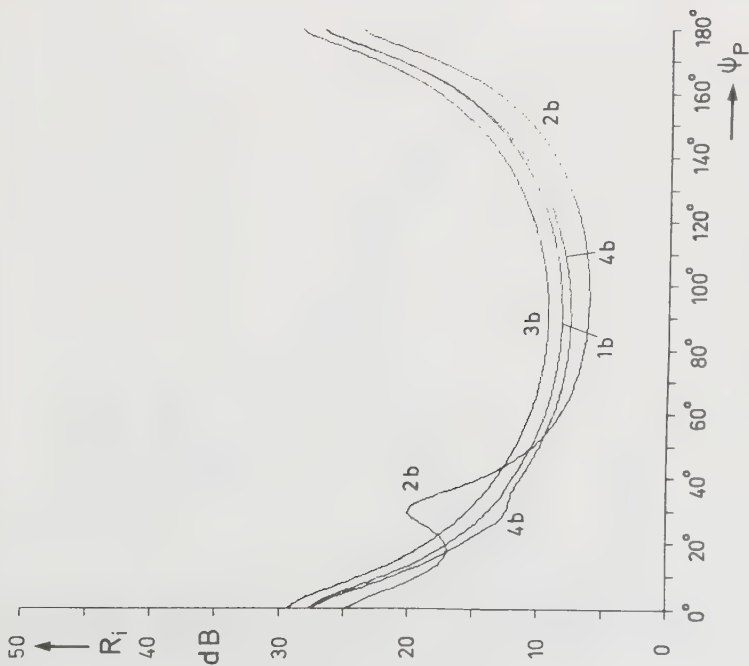


Bild 5.10: Pol-Nullstellen-Test
Struktur 1,2,3,4
 $\rho_P = 0.9$
 $\psi_N = \pm 30^\circ$

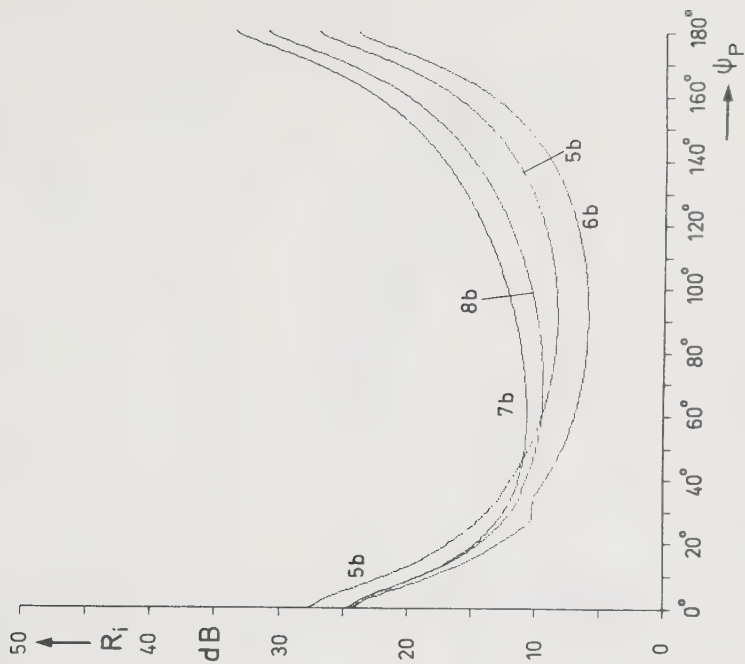


Bild 5.11: Pol-Nullstellen-Test
Struktur 5,6,7,8
 $\rho_P = 0.9$
 $\psi_N = \pm 30^\circ$

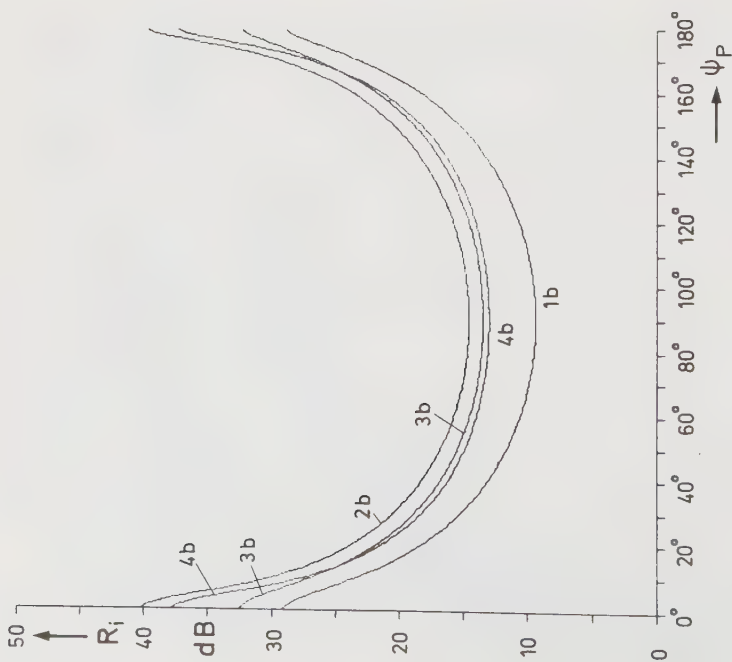


Bild 5.12: Allpaß-Test $\rho_P = 0.9$
Struktur 1, 2, 3, 4

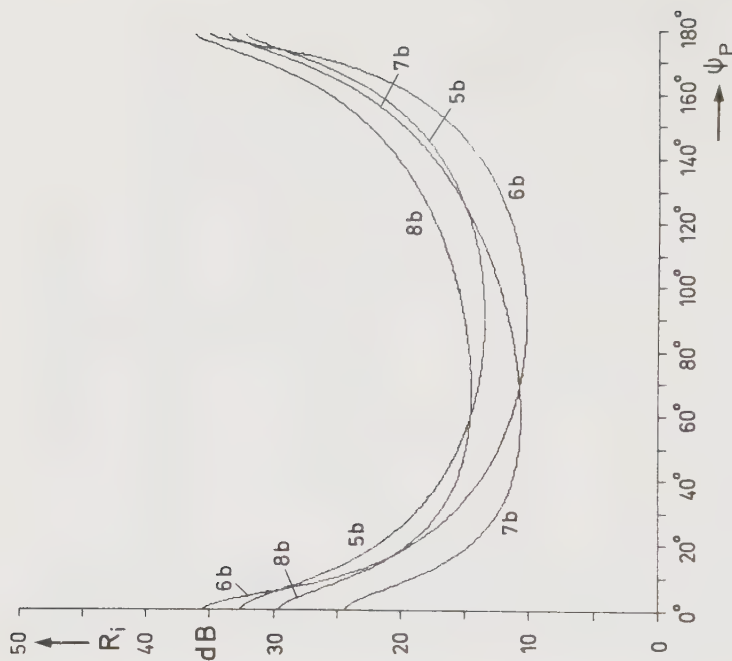


Bild 5.13: Allpaß-Test
Struktur 5, 6, 7, 8 $\rho_P = 0.9$

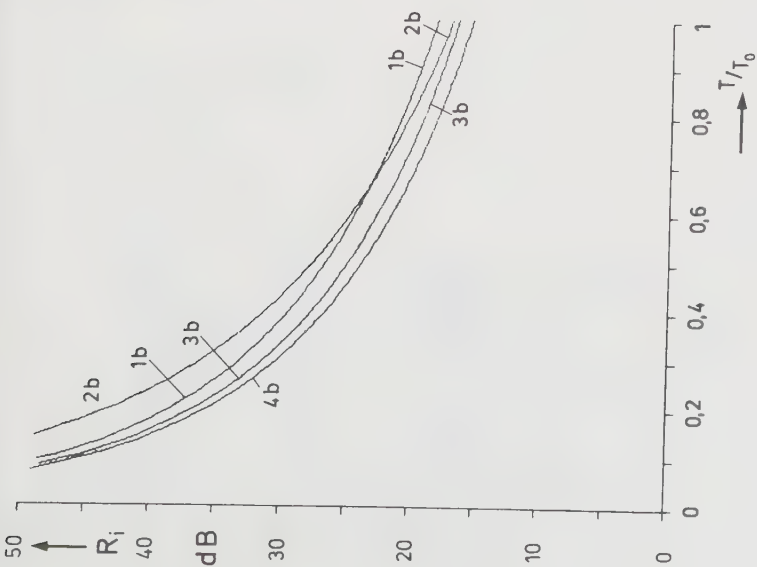


Bild 5.14: Abtastintervall-Test
Struktur 1,2,3,4

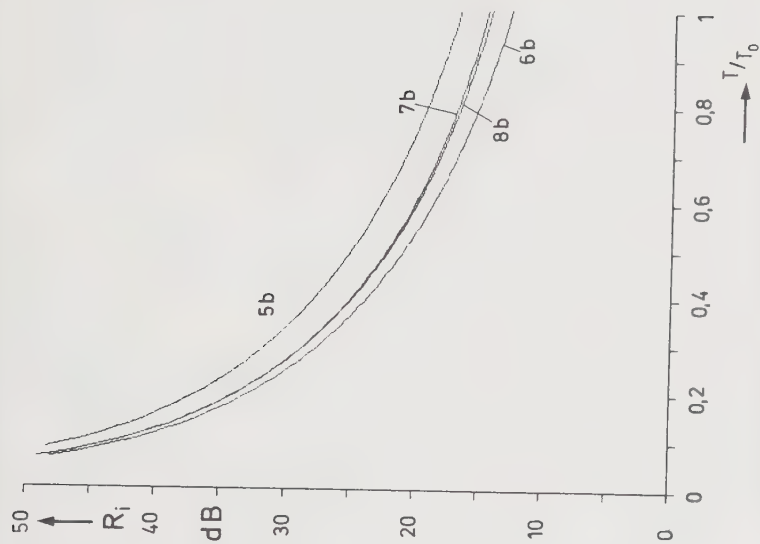


Bild 5.15: Abtastintervall-Test
Struktur 5,6,7,8

Koeffizienten unskaliert:

$$\begin{aligned} d_2 &= b_2 \\ d_1 &= b_1 - b_2 c_1 \\ d_0 &= b_0 - b_2 c_0 \end{aligned} \quad (5.27)$$

Skalierung:

$$\begin{aligned} {}_5F_3(z) &= (d_1 z + d_0) / C(z) \\ {}_5F_4(z) &= {}_5H(z) \end{aligned} \quad (5.28)$$

$$\begin{aligned} b_2' &= b_2 \left[\max_i \{ \| {}_5F_i \|_\infty \} \right]^{-1} \quad i = 3, 4 \\ b_3 &= 1 ; \quad b_4 = b_2 / b_2' \\ d_i' &= d_i / b_4, \quad i = 0, 1, 2 \end{aligned} \quad (5.29)$$

Fehlerübertragungsfunktionen:

$$\begin{aligned} {}_5G_1(z) &= b_4 / C(z) \\ {}_5G_2(z) &= b_4 z / C(z); \quad {}_5G_3(z) = b_4 \end{aligned} \quad (5.30)$$

Pol-Nullstellen-Test: Bild 5.11
Allpaß-Test : Bild 5.13
Abtastintervall-Test: Bild 5.15

f) Struktur 6 - 1C nach Jackson

Signalflußgraph: Tabelle 5.2

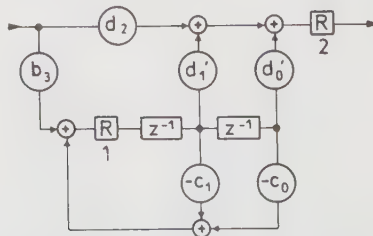


Bild 5.16: Blockschaltbild Struktur 6 b

Übertragungsfunktion: ${}_6H(z) = {}_5H(z)$ nach Gl. (5.26)

Koeffizienten unskaliert: siehe Gl. (5.27)

Skalierung: ${}_6F_1(z) = z^2/C(z)$ (5.31)

$$b_3 = ||{}_6F_1||_{\infty}^{-1}$$

$$b_2' = b_2/b_3 ; b_4 = 1$$

$$d_i' = d_i/b_3 ; i = 0, 1$$
 (5.32)

Fehlerübertragungsfunktionen:

$${}_6G_1(z) = (d_1'z + d_0')/C(z)$$

$${}_6G_2(z) = 1$$
 (5.33)

Pol-Nullstellen-Test: Bild 5.11

Allpaß-Test : Bild 5.13

Abtastintervall-Test: Bild 5.15

g) Struktur 7 - S_{b2}^T nach Parker und Hess

Signalflußgraph: Tabelle 5.2

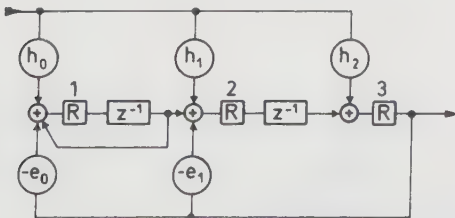


Bild 5.17: Blockschaltbild Struktur 7b

Übertragungsfunktion:

$${}_7H(z) = \frac{h_2 z^2 + (h_1 - h_2) z + (h_0 - h_1)}{z^2 + (e_1 - 1) z + (e_0 - e_1)} \quad (5.34)$$

Koeffizienten unskaliert:

$$\begin{aligned} h_2 &= b_2 \\ h_1 &= b_2 + b_1 & e_1 &= 1 + c_1 \\ h_0 &= b_2 + b_1 + b_0 & e_0 &= 1 + c_1 + c_0 \end{aligned} \quad (5.35)$$

Skalierung: ${}_7F_3(z) = {}_7H(z) \quad (5.36)$

$$b_3 = 1 ; b_2' = b_2 ; b_4 = 1 \quad (5.37)$$

Fehlerübertragungsfunktionen:

$$\begin{aligned} {}_7G_1(z) &= 1/C(z) \\ {}_7G_2(z) &= (z-1)/C(z) \\ {}_7G_3(z) &= (z^2 - z)/C(z) \end{aligned} \quad (5.38)$$

Pol-Nullstellen-Test: Bild 5.11

Allpaß-Test : Bild 5.13

Abtastintervall-Test: Bild 5.15

h) Struktur 8 - $S_{b_2}^T$ nach Parker und Hess

Signalflußgraph: Tabelle 5.2

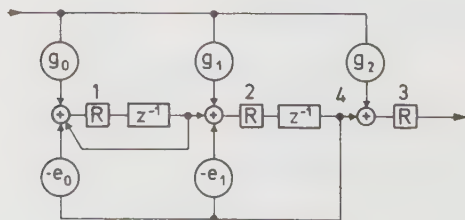


Bild 5.18: Blockschaltbild - Struktur 8 b

Übertragungsfunktion:

$${}_8H(z) = \frac{g_2 z^2 + (g_1 + g_2 e_1 - g_2) z + (g_0 - g_1 + g_2 (e_0 - e_1))}{z^2 + (e_1 - 1) z + (e_0 - e_1)} \quad (5.39)$$

Koeffizienten unskaliert:

$$\begin{aligned} g_2 &= b_2 \\ g_1 &= b_1 - b_2 c_1 ; \quad e_1 = 1 + c_1 \\ g_0 &= b_0 + b_1 - b_2 (c_0 + c_1) ; \quad e_0 = 1 + c_1 + c_0 \end{aligned} \quad (5.40)$$

Skalierung:

$${}_8F_4(z) = (g_1 z + (g_0 - g_1)) / C(z) \quad (5.41)$$

$${}_8F_3(z) = {}_8H(z)$$

$$b_3 = 1$$

$$b_2' = b_2 \left[\max_i \{ \| {}_5F_i \|_\infty \} \right]^{-1}, \quad i = 3, 4$$

$$b_4 = b_2 / b_2'$$

$$g_i' = g_i / b_4, \quad i = 0, 1, 2$$

Fehlerübertragungsfunktion:

$${}_8G_1(z) = b_4 / C(z) \quad (5.43)$$

$${}_8G_2(z) = b_4 (z - 1) / C(z)$$

$${}_8G_3(z) = b_4$$

Pol-Nullstellen-Test: Bild 5.11

Allpaß-Test : Bild 5.13

Abtastintervall-Test: Bild 5.15

i) Struktur 14 - 6. kanonische Form nach Hwang

Signalflußgraph: Tabelle 5.5

Übertragungsfunktion: siehe Tabelle 5.6

Skalierung, Fehlerübertragungsfunktion siehe i)

Pol-Nullstellen-Test: Bild 5.23

m) Struktur 17 - Brückenstruktur nach Markel und Gray

Signalflußgraph: Tabelle 5.6

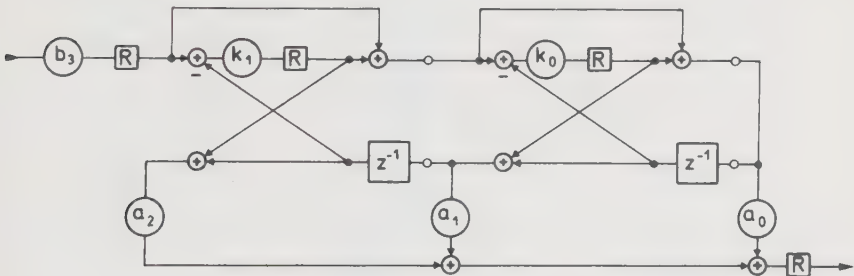


Bild 5.21: Blockschaltbild Struktur 17 b

Übertragungsfunktionen: siehe Tabelle 5.6

Skalierung, Fehlerübertragungsfunktionen siehe i)

Pol-Nullstellen-Test: Bild 5.23

5.2.3.2 Vergleich der Ergebnisse

Der Vergleich der Ergebnisse des Pol-Nullstellentests der Strukturen 1-8 zeigt, daß die Struktur 6 das günstigste Verhalten erbringt. In einen weiten Bereich - für Pol- und Nullstelle nicht zu nahe beieinander liegend - gleichen die Ergebnisse der Struktur 2. Jedoch weist diese Struktur für Pole in der Nähe der Nullstellen aufgrund der ungünstigen Aussteuerung einen Anstieg der Rauschleistung auf. In Bezug auf den Allpaß-test wird die Struktur 6 nur von der Struktur 1 (der 1. kan. Form) geringfügig übertriften. Der Abtastintervall-Test liefert ebenfalls für die Struktur 6 die günstigsten Ergebnisse. Aus

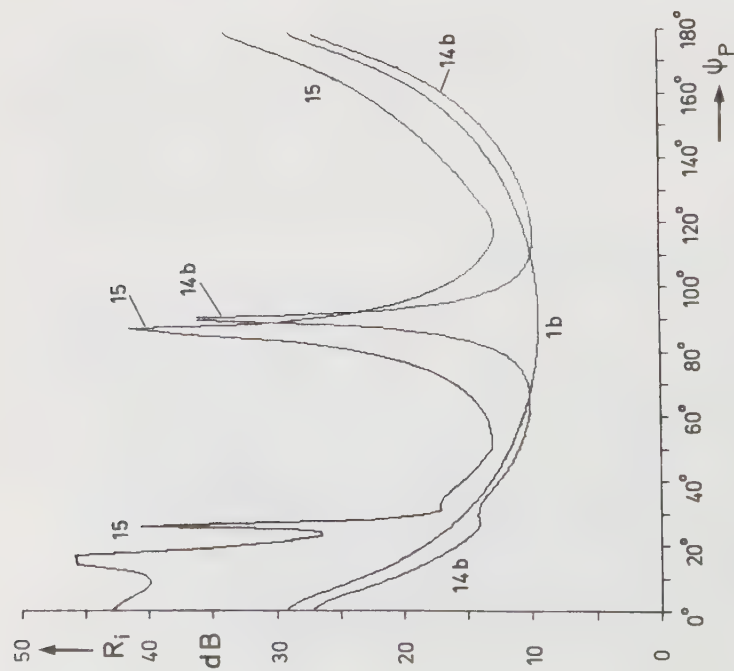


Bild 5.22: Pol-Nullstellen-Test
Struktur 1, 14, 15

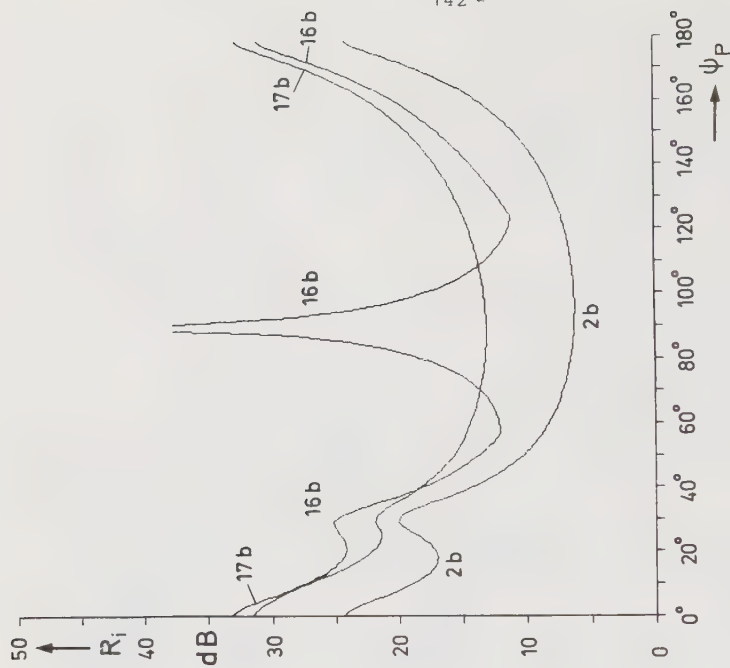


Bild 5.23: Pol-Nullstellen-Test
Struktur 2, 16, 17

der Sicht dieser drei Tests erscheint die Struktur 6 eine günstige Lösung für einen Block 2. Grades zu sein. Nachteilig wirkt sich jedoch für diese Struktur aus, daß im Gegensatz zur 1. und 2. kanonischen Form bei Nullstellen auf dem Einheitskreis ($b_2 = b_0$) kein Multiplizierer eingespart werden kann. Ein weiterer Vorteil der 2. kan. Form ist die mögliche Reduzierung der Rauschquellen pro Block bei der Kaskadierung.

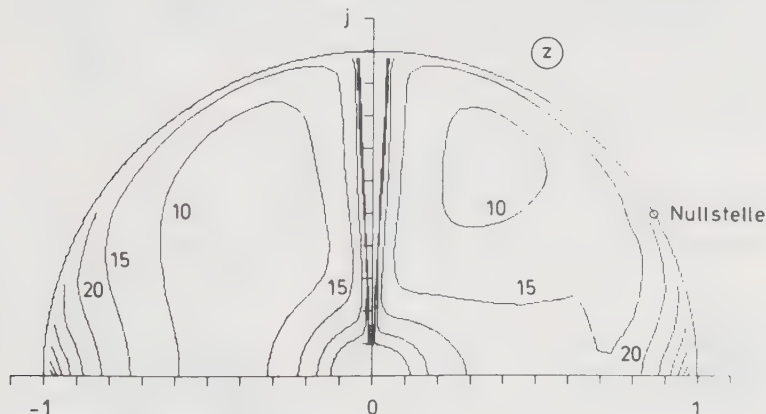


Bild 5.24: Linien gleicher Rauschzahlen Struktur 14

$$\psi_N = \pm 30^\circ \quad \rho_N = 1$$

Die Strukturen 1, 2, 5, 6, 7, 8 besitzen die gleiche Koeffizientenempfindlichkeit $|77|$. Dennoch ist zu berücksichtigen, daß bei den Strukturen 5, 6, 7 und 8 auch bei $b_2 = 1$ (z.B. Realisierung der Kaskadenschaltung mit Vorfaktoren eine Vergrößerung des Koeffizientenzahlenbereichs auftreten kann und somit nach Abschnitt 2.1 bei gleicher Gesamtkoeffizientenwortlänge ein zusätzlicher Koeffizientenfehler entstehen kann.

Stellt man die zusätzliche Bedingung, daß die Zähler und die Nennerkoeffizienten voneinander unabhängig sein sollen, so erweist sich die 2. kanonische Form (Struktur 2) im allgemeinen in Bezug auf das Rundungsrauschen als am günstigsten.

Den Strukturen 14, 15, 16, 17 ist gemeinsam, daß ein Teil der Multiplizierer-Koeffizienten durch Division der ursprünglichen

Parameter entstehen. Diese Koeffizienten können daher über alle Grenzen anwachsen. Die Strukturen sind daher nicht für alle Pol-Nullstellenlagen realisierbar. Gleichzeitig entstehen hierbei Bereiche mit extrem ungünstigen Aussteuerbedingungen. Entsprechend steigen die Rauschzahlen in den Bildern 5.22 und 5.23 an. Die Einschränkungen der möglichen Pollagen bei vorgegebener Nullstelle verdeutlichen darüber hinaus die beiden Diagramme (Bild 5.24 und 5.25) mit Linien gleicher Rauschzahlen.

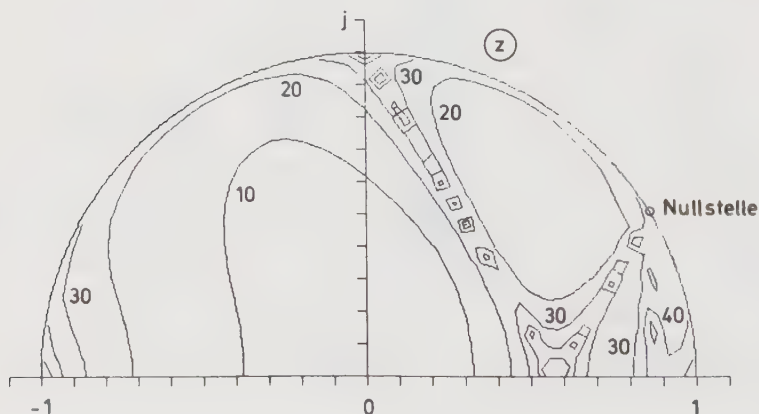


Bild 5.25: Linien gleicher Rauschzahlen Struktur 15

$$\psi_N = \pm 30^\circ \quad \rho_N = 1$$

5.2.4 Rauschverhalten bei unterschiedlicher Arithmetik

Der Einfluß verschiedener Rundungsarithmetiken und die Zusammenfassung der ungerundeten Teilprodukte auf das Rauschverhalten der Strukturen soll am Beispiel der ersten (Struktur 1) und der zweiten (Struktur 2) kanonischen Form nach [1] untersucht werden. Die Skalierung der Strukturen wird entsprechend Abschnitt 5.2.3.1 a,b vorgenommen. Die Beurteilung erfolgt nach dem Pol-Nullstellen-Test.

5.2.4.1 Runden

Die Blockschaltbilder erklären die verwendete Arithmetik.

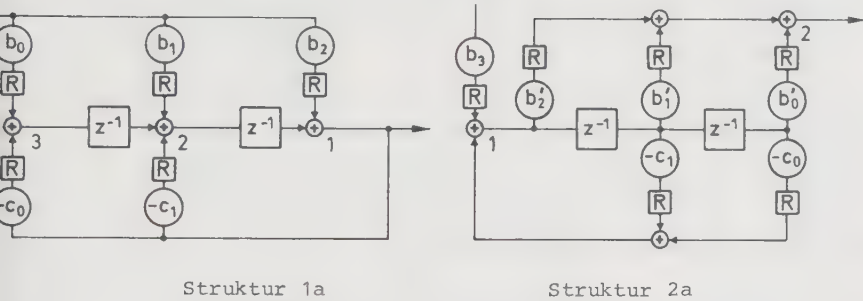


Bild 5.26: Runden nach jeder Multiplikation

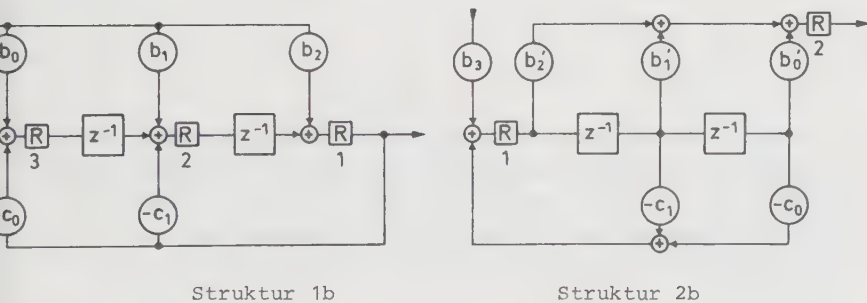


Bild 5.27: Runden nach Zusammenfassung der Teilprodukte

Liegen die Nullstellen der Übertragungsfunktion bei den Strukturen 1a und 2a auf dem Einheitskreis ($b_2 = b_0$), so entstehen bei der Rundung unmittelbar nach den Multiplizierern mit b_2 und b_0 exakt die gleichen Fehler, die am Ausgang der Schaltung mit einer Verzögerung von $2T$ zueinander addiert werden [7]. Durch entsprechende Ersatzschaltungen nach Bild 5.28 können die Korrelationen der Fehler berücksichtigt werden.

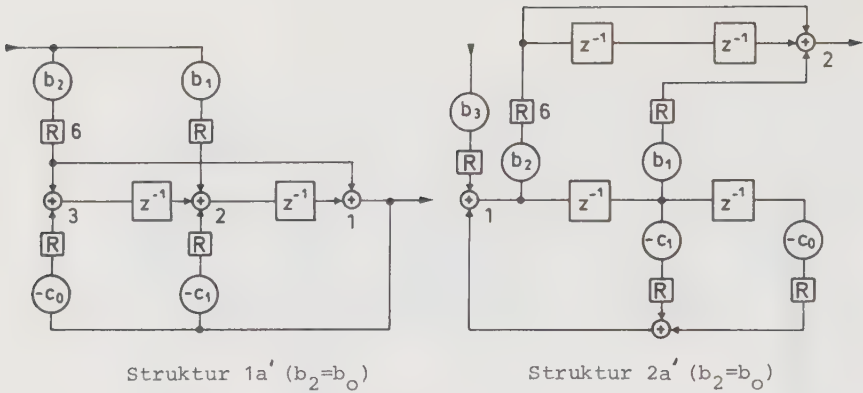


Bild 5.28: Ersatzschaltungen für $b_2=b_0$

In Ergänzung zu Gl. (5.13) und Gl. (5.17) ergeben sich die neuen Fehlerübertragungsfunktionen zu

$${}_1G_6(z) = (z^2+1)/C(z) \quad (5.44)$$

und

$${}_2G_6(z) = z^2+1. \quad (5.45)$$

Sind die zu realisierenden Übertragungsfunktionen auf Nullstellen auf den Einheitskreis beschränkt, so kann bei der ersten kanonischen Form die Ersatzschaltung direkt verwendet werden. Bei der zweiten kanonischen Form ist eine Verschiebung der Multiplikation entsprechend Bild 5.29 notwendig.

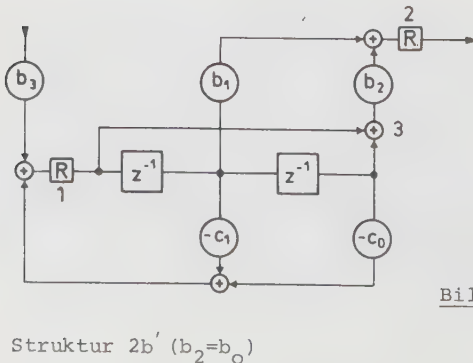


Bild 5.29: 2. kanonische Form für $b_2=b_0$

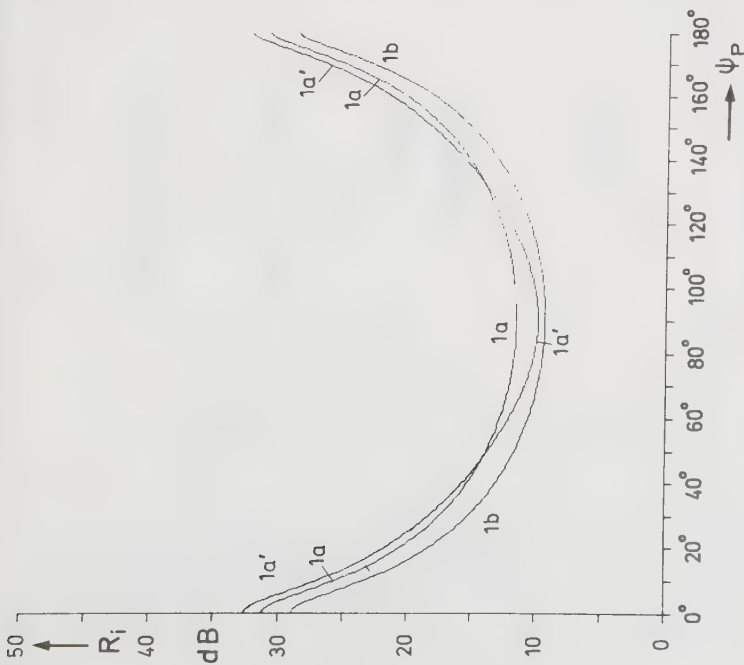


Bild 5.30: Runden bei untersch. Zusammenfassung der Teilprodukte
Struktur 1

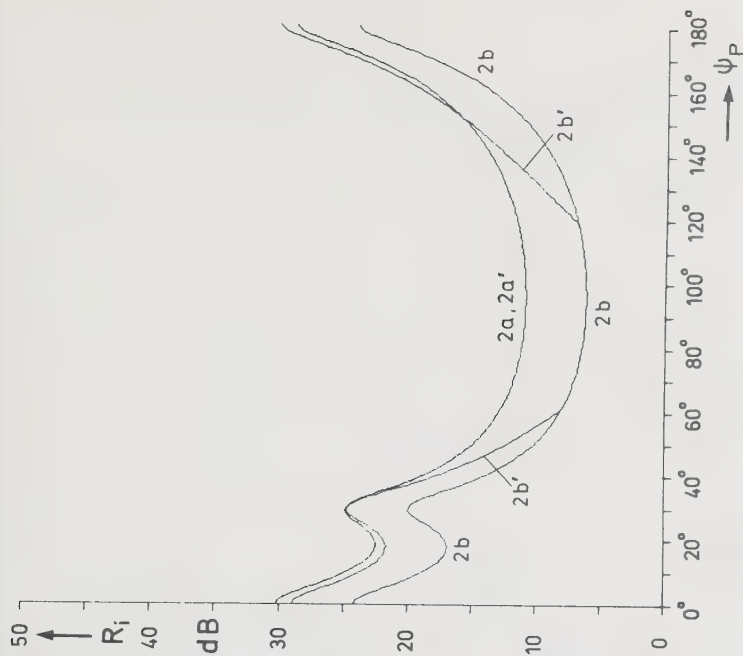


Bild 5.31: Runden bei untersch. Zusammenfassung der Teilprodukte
Struktur 2

Zur Vermeidung von Übersteuerungen muß dann zusätzlich der Eingang des Multiplizierers b_2 überprüft werden. Mit Gl. (5.10) erhält man in Ergänzung zu Gl. (5.14), (5.15)

$$2^{F_6}(z) = (z^2+1)/C(z) , \quad (5.46a)$$

und

$$b_3 = \min_i \{ ||2^{F_i}||_{\infty}^{-1} \} , \quad i = 1, 6 . \quad (5.46b)$$

Ein Vergleich der Kurven in Bild 5.30 und 5.31 zeigt die Vorteile der Zusammenfassung der ungerundeten Teilprodukte. Dies wird besonders bei der 2. kanonischen Form deutlich, da hier die von der Nennerfunktion $C(z)$ verstärkten Rauschquellen auf eine reduziert werden können.

5.2.4.2 Zweierkomplement-Abschneiden

Die einfachste Art einer Verkürzung ist das Zweier-Komplement-Abschneiden nach Abschnitt 4.2.1.2. Mit Gl. (4.78) kann der zum Rauschen zusätzliche Gleichanteil berechnet und so bei der Beurteilung berücksichtigt werden.

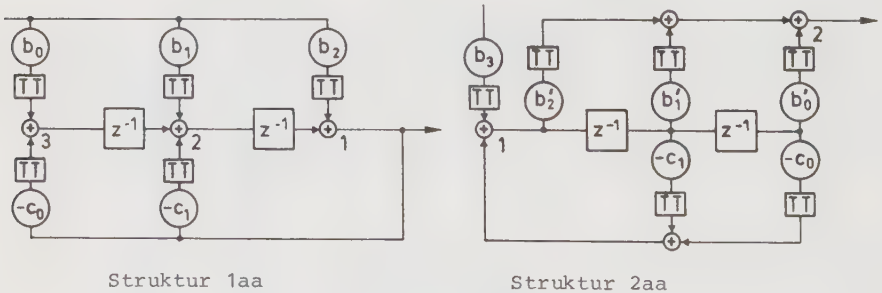


Bild 5.32: Zweierkomplement-Abschneiden nach dem Multiplizieren.

Die Ergebnisse des Pol- Nullstellentests zeigen die Bilder 5.33 und 5.34. Die entsprechenden Werte bei Runden (Strukturen 1a und 1b) sind zum Vergleich eingetragen. Ein deutlicher Anstieg der Rauschleistung bei Zweierkomplement-Abschneiden ist für $\Psi_P \rightarrow 0$ zu erkennen.

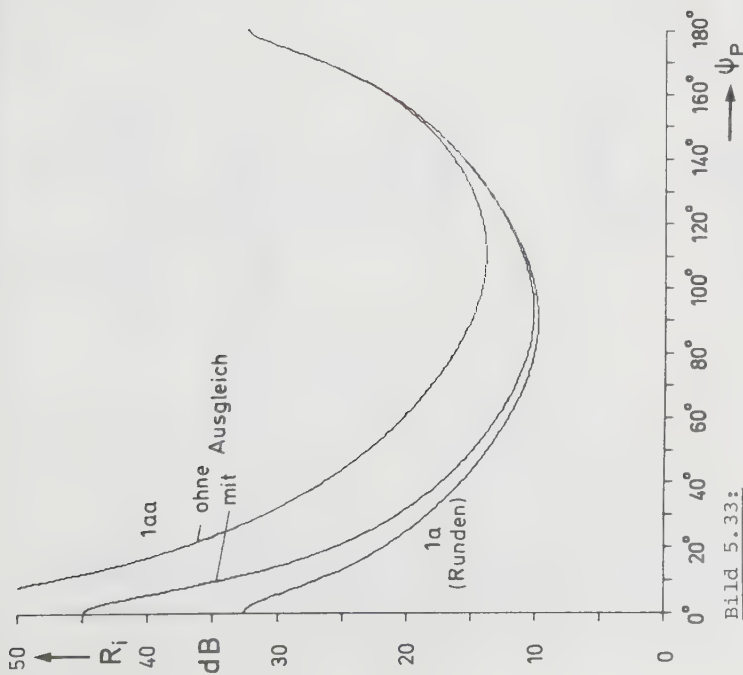


Bild 5.33:

Pol-Nullstellen-Test $\psi_N = \pm 30^\circ$, $\rho_P = 0.9$
Struktur 1 aa, Zweierkomplement-Abschneiden

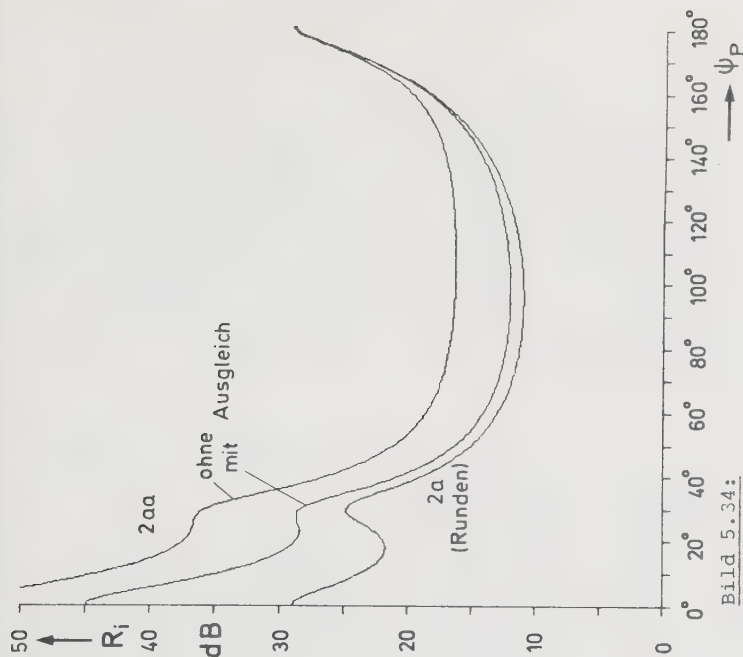


Bild 5.34:

Pol-Nullstellen-Test $\psi_N = \pm 30^\circ$, $\rho_P = 0.9$
Struktur 2 aa, Zweierkomplement-Abschneiden

Der im Fehlersignal enthaltene Gleichanteil entsteht unabhängig vom Signal. Ein teilweiser Ausgleich ist möglich, wenn die Gleichanteile der einzelnen Fehler mit unterschiedlichem Vorzeichen zum Ausgang hin übertragen werden. Bei negativen Koeffizienten kann durch Komplementieren nach der Nichtlinearität eine negative Fehlerübertragungsfunktion erzeugt werden. Im vorliegenden Beispiel wurde der Ausgleich für je einen Zähler- und Nennerkoeffizienten durchgeführt. Eine deutliche Verbesserung wird sichtbar.

5.2.4.3 Betragsmäßiges Abschneiden

Durch betragsmäßiges Abschneiden im rekursiven Teil der Schaltung kann bei der 1. und 2. kanonischen Form eine weitgehende Unterdrückung von Grenzyklen erreicht werden [61], [65], [66]. Der Einfluß dieser Maßnahmen auf das Rauschen soll hier näher untersucht werden. In den Bildern 5.35 bis 5.37 werden verschiedene Strukturen gezeigt.

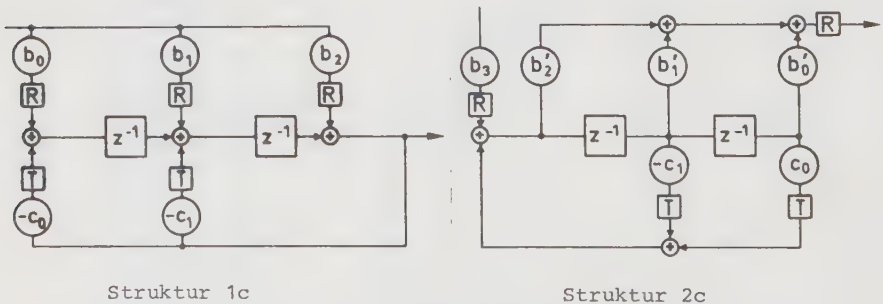
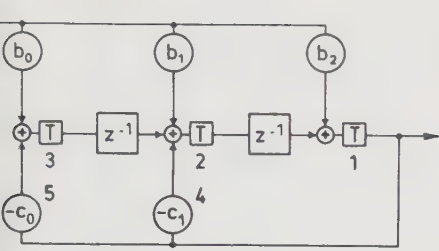
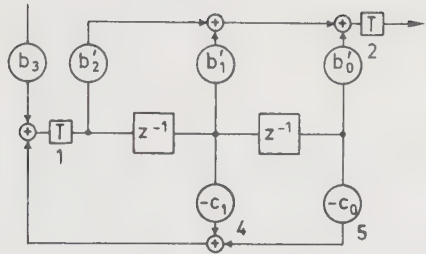


Bild 5.35: Getrenntes betragsmäßiges Abschneiden nach den Nenner-Multiplikationen

Zur Berechnung der Rauschleistung bei betragsmäßigem Abschneiden werden zusätzlich die Übertragungsfunktionen vom Eingang zu den Nichtlinearitäten benötigt. Mit Gl. (5.10) erhält man

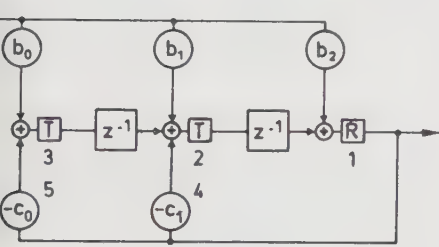


Struktur 1d

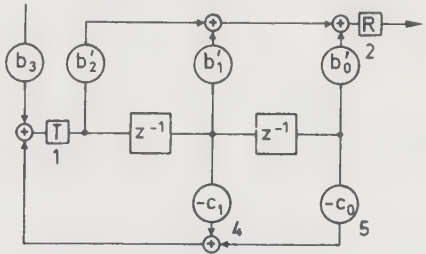


Struktur 2d

Bild 5.36: Betragsmäßiges Abschneiden nach Zusammenfassung der Teilprodukte



Struktur 1e



Struktur 2e

Bild 5.37: Betragsmäßiges Abschneiden nach zusammengefaßten Nennerprodukten

Struktur 1

$$\begin{aligned}
 {}_1F_1(z) &= H(z) \\
 {}_1F_2(z) &= ((b_1 - b_2 c_1) z^2 + (b_0 - b_2 c_0) z) / C(z) \\
 {}_1F_3(z) &= ((b_0 - b_2 c_0) z^2 + (b_0 c_1 - b_1 c_0) z) / C(z) \\
 {}_1F_4(z) &= -c_1 H(z) \\
 {}_1F_5(z) &= -c_0 H(z)
 \end{aligned} \tag{5.47}$$

Struktur 2

$$\begin{aligned}
 {}_2F_1'(z) &= b_3 z^2 / C(z) \quad b_3 \text{ nach Gl. (5.15) bzw. (5.46b) (5.48)} \\
 {}_2F_2'(z) &= {}_2F_2(z) = H(z)
 \end{aligned}$$

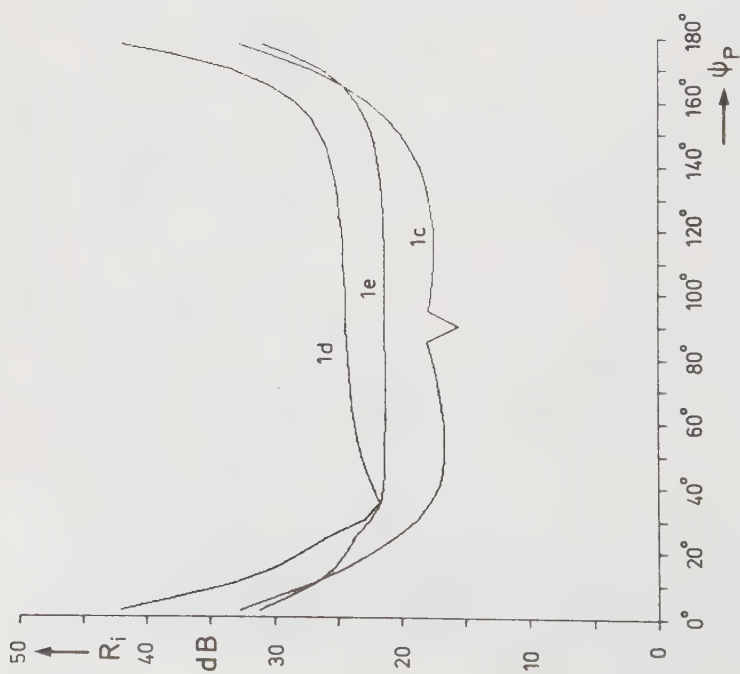


Bild 5.38: Pol-Nullstellen-Test $\psi_N = \pm 30^\circ$ $\rho_P = 0.9$
 1. kanonische Form mit
 betr. Abschnneiden an versch. Stellen

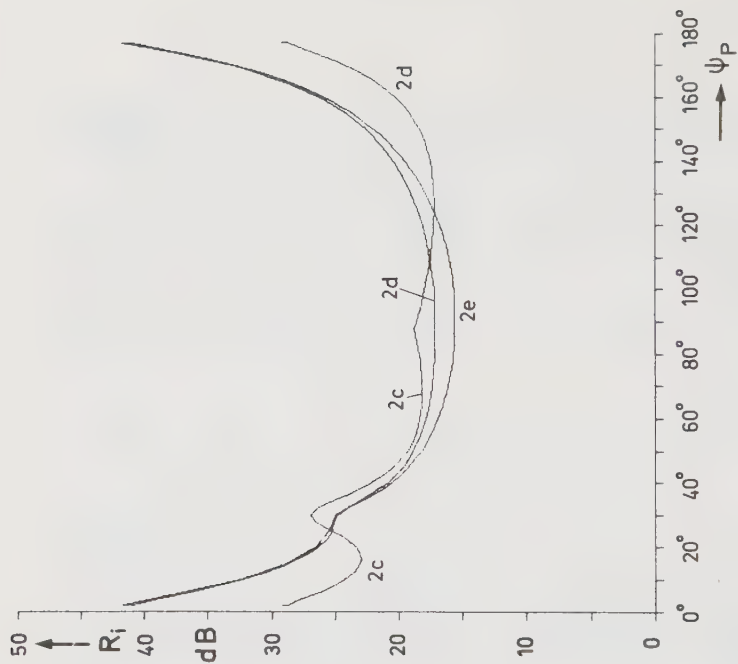


Bild 5.39: Pol-Nullstellen-Test $\psi_N = \pm 30^\circ$ $\rho_P = 0.9$
 2. kanonische Form mit
 betr. Abschnneiden an versch. Stellen

$${}_2F_4'(z) = -b_3 c_1 z / C(z)$$

$${}_2F_5'(z) = -b_3 c_0 / C(z)$$

Der Pol-Nullstellen-Test (Bild 5.38 u. 5.39) wird mit Hilfe des Signum-Modells nach Gl. (4.66) durchgeführt. Eine nähere Betrachtung der Kurven zeigt, daß eine Zusammenfassung ungerundeter Teilprodukte im Gegensatz zum Runden hier nicht unbedingt zu günstigeren Ergebnissen führen muß. Speziell für die Struktur 1c wurde in [7] gezeigt, daß sich in bestimmten Koeffizientenbereichen die mit dem Signal korrelierten Fehleranteile zweier verschiedener Quellen gegenseitig aufheben können. Ein Vergleich mit Runden (Bild 5.30) zeigt einen generellen Anstieg der Rauschleistung bei betragsmäßigem Abschneiden.

5.2.5 Rauschverhalten grenzzyklenfreier Strukturen

Bei den in Tabelle 5.4 angegebenen Strukturen kann das Entstehen von Grenzzyklen vermieden werden. Das Rauschverhalten dieser Strukturen wird im folgenden näher untersucht.

a) Struktur_11 nach Meerkötter und Wegener [62]

Signalflußgraph: Tabelle 5.4

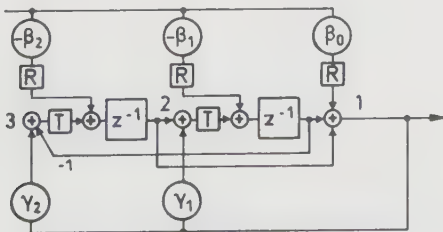


Bild 5.40: Blockschaltbild Struktur 11

Übertragungsfunktion:

$${}_{11}H(z) = \frac{\beta_0 z^2 + (\beta_1 + \beta_2)z + (\beta_0 + \beta_1 - \beta_2)}{z^2 + (\gamma_1 + \gamma_2)z + (1 + \gamma_1 - \gamma_2)} = \frac{B(z)}{C(z)} \quad (5.49)$$

Koeffizienten unskaliert:

$$\begin{aligned} \beta_0 &= b_2 \\ \beta_1 &= (b_0 - b_1 - b_2)/2 & \gamma_1 &= (c_0 - c_1 - 1)/2 \\ \beta_2 &= (b_2 - b_1 - b_0)/2 & \gamma_2 &= (1 - c_1 - c_0)/2 \end{aligned} \quad (5.50)$$

Teilübertragungsfunktionen:

$$\begin{aligned} {}_{11}F_1(z) &= {}_{11}H(z) \\ {}_{11}F_2(z) &= \frac{\beta_0 \gamma_1 z^2 + (\beta_0 \gamma_2 - \beta_1 \gamma_1 - \beta_2 (1 + \gamma_1))z + \beta_1 (1 + \gamma_1 - \gamma_2)}{C(z)} \\ {}_{11}F_3(z) &= \frac{\beta_0 \gamma_2 z^2 - (\beta_0 \gamma_1 + \beta_1 (\gamma_2 - 1) + \beta_2 \gamma_2)z + \beta_2 (1 + \gamma_1 - \gamma_2)}{C(z)} \end{aligned}$$

$$\text{Skalierung:} \quad b_3 = 1 ; \quad b_2' = b_2 ; \quad b_4 = 1 \quad (5.52)$$

Fehlerübertragungsfunktionen:

$$\begin{aligned} {}_{11}G_1(z) &= (z^2 + 1)/C(z) \\ {}_{11}G_2(z) &= (z - 1)/C(z) \\ {}_{11}G_3(z) &= (z + 1)/C(z) \\ {}_{11}G_4(z) &= 2z/C(z) \quad \text{für } \beta_1 = \beta_2 \text{ korrelierte Rauschquellen} \end{aligned} \quad (5.53)$$

Pol-Nullstellen-Test: Bild 5.43

b) Struktur_12 - Transponierte zu Struktur 11

Signalflußgraph: Tabelle 5.4

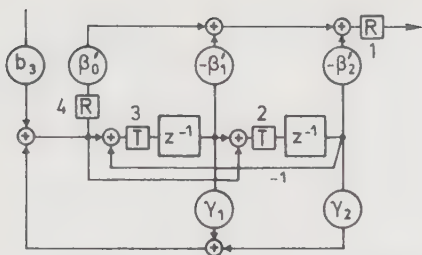


Bild 5.41: Blockschaltbild Struktur 12

Übertragungsfunktion: ${}_{12}H(z) = {}_{11}H(z)$ nach Gl. (5.49)

Koeffizienten unskaliert: siehe Gl. (5.50)

Teilübertragungsfunktionen:

$$\begin{aligned} {}_{12}F_1(z) &= {}_{12}H(z) \\ {}_{12}F_2(z) &= (z^2 + z)/C(z) \\ {}_{12}F_3(z) &= (z^2 - z)/C(z) \\ {}_{12}F_4(z) &= (z^2 + 1)/C(z) \end{aligned} \quad (5.54)$$

Skalierung:

$$b_3 = \left[\max \{ || {}_{12}F_i ||_{\infty} \} \right]^{-1} \quad i = 2, 3, 4 \quad (5.55)$$

$$b_2' = b_2/b_3 ; \quad b_4 = 1$$

$$\beta_i' = \beta_i/b_3 ; \quad i = 0, 1, 2 \quad (5.56)$$

Fehlerübertragungsfunktionen:

$$\begin{aligned} {}_{12}G_1(z) &= 1 \\ {}_{12}G_2(z) &= \frac{(\beta_0' \gamma_2 - \beta_2')z + (\gamma_1(\beta_2' - \beta_0') + \beta_1'(1 - \gamma_2))}{C(z)} \\ {}_{12}G_3(z) &= \frac{(\beta_0' \gamma_1 - \beta_1')z + (\gamma_2(\beta_0' + \beta_1') - \beta_2'(1 + \gamma_1))}{C(z)} \\ {}_{12}G_4(z) &= \beta_0' \end{aligned} \quad (5.57)$$

Pol-Nullstellen-Test: Bild 5.43

c) Struktur_13 - nach Verkroost und Butterweck

Signalflußgraph: Tabelle 5.4

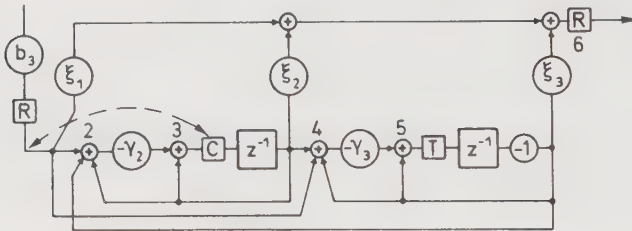


Bild 5.42: Blockschaltbild Struktur 13

Übertragungsfunktion:

$${}_{13}^H(z) = \frac{\xi_1 z^2 + 2((\gamma_2 + \gamma_3 - 1)\xi_1 - \xi_2 \gamma_2 - \gamma_3 \xi_3)z + 2(\xi_1(\gamma_2 - \gamma_3) - \gamma_2 \xi_2 + \gamma_3 \xi_3)}{z^2 + (\gamma_2 - \gamma_3)z + (\gamma_2 + \gamma_3 - 1)} \quad (5.58)$$

Koeffizienten unskaliert:

$$\begin{aligned} \xi_1 &= b_2 \\ \xi_2 &= \frac{(c_0 + c_1)b_2 - b_1 - b_0}{1 + c_1 + c_0} ; \quad \gamma_2 = (1 + c_1 + c_0)/2 \\ \xi_3 &= \frac{(c_0 - c_1)b_2 + b_1 - b_0}{1 - c_1 + c_0} ; \quad \gamma_3 = (1 - c_1 + c_0)/2 \end{aligned} \quad (5.59)$$

Teilübertragungsfunktionen:

$$\begin{aligned} {}_{13}^{F_2}(z) &= (z^2 - 1)/C(z) \\ {}_{13}^{F_3}(z) &= -\gamma_2(z^2 + z)/C(z) \\ {}_{13}^{F_4}(z) &= {}_{13}^{F_2}(z) \end{aligned} \quad (5.60)$$

$${}_{13}F_5(z) = -\gamma_3(z^2 - z)/C(z)$$

Skalierung:

$$b_3 = \left[\max_i \{ ||{}_{13}F_i||, 1 \} \right]^{-1} \quad i=2,3,4,5$$

$$b_2' = b_2/b_3 \quad b_4 = 1 \quad (5.61)$$

$$\xi_i' = \xi_i/b_3 \quad i=1,2,3 \quad (5.62)$$

Hilfsübertragungsfunktion für "kontrolliertes Runden" (Abschnitt 4.2.1.4):

Mit dem negativen Eingangssignal als Referenz erhält man nach Gl. (4.31)

$${}_{13}F_a(z) = {}_{13}F_3(z) + 1 \quad (5.63)$$

Fehlerübertragungsfunktionen:

$${}_{13}G_1(z) = {}_{13}H(z)/b_3$$

$${}_{13}G_3(z) = (\xi_2'z + (\xi_2'(1-\gamma_3) + \xi_3'\gamma_3))/C(z) \quad (5.64)$$

$${}_{13}G_5(z) = (-\xi_3'z + (\xi_3'(1-\gamma_2) + \xi_2'\gamma_2))/C(z)$$

$${}_{13}G_6(z) = 1$$

Pol-Nullstellen-Test: Bild 5.43

Die Pol-Nullstellen-Tests für die drei grenzzyklenfreien Strukturen sind im Bild 5.43 dargestellt. Zum Vergleich werden diese Strukturen mit Runden betrieben. Man sieht, daß die Forderung nach Grenzzyklenfreiheit das Rauschen erheblich ansteigen läßt. Dies wird besonders im Vergleich mit den Strukturen 2 b (Bild 5.10) und 6 b (Bild 5.11) deutlich.

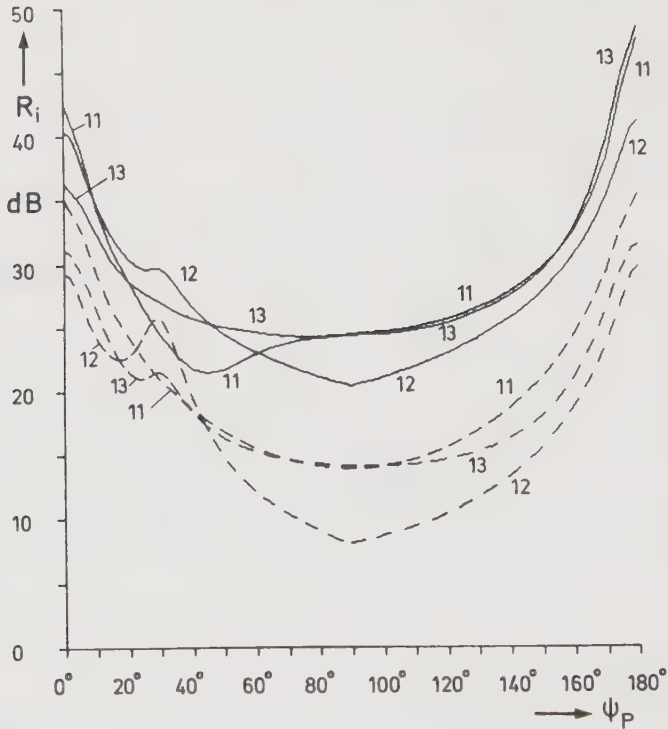


Bild 5.43: Pol-Nullstellen-Test der Strukturen 11, 12, 13
 — , zum Vergleich mit Runden ----

Entsprechend den Ausführungen in [7] und [58] entsteht bei der Verkürzung der Zustandsvariablen mit betragsmäßigem Abschneiden ein mit dem Signal linear korrelierter Fehleranteil. Wie man aus Gl. (4.21) sieht, hängt die Amplitude des gesamten mit dem Signal korrelierten Fehleranteils nur von der Quantisierungsstufe Q und nicht von der Aussteuerung des Signals ab. Somit ist auch der mit dem Signal linear korrelierte Fehleranteil von der Amplitude des Signals unabhängig.

Eine Einbeziehung des linear korrelierten Fehlers in die Übertragungsfunktion zur Reduzierung des Gesamtfehlers erscheint im allgemeinen jedoch nicht sinnvoll, da der äquivalente Verstärkungsfaktor nach [7] von der Aussteuerung des Signals selbst abhängt.

5.2.6 Das Rauschverhalten von Strukturen mit rückgekoppelten Speichern und inneren Skalierern

a) Struktur 9 - nach Agarwal und Burrus [41]

Signalflußgraph: Tabelle 5.3

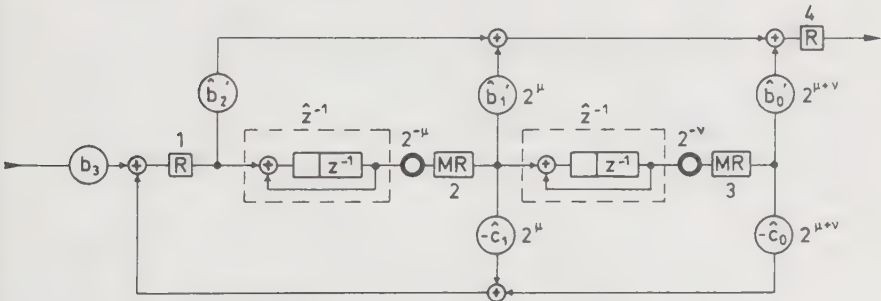


Bild 5.44: Blockschaltbild Struktur 9

Übertragungsfunktion:
$$\frac{\hat{b}_2 z^2 + (\hat{b}_1 - 2\hat{b}_2) z + (\hat{b}_0 - \hat{b}_1 + \hat{b}_2)}{z^2 + (\hat{c}_1 - 2) z + (\hat{c}_0 - \hat{c}_1 + 1)} \quad (5.65)$$

Koeffizienten unskaliert:

$$\begin{aligned} \hat{b}_2 &= b_2 \\ \hat{b}_1 &= 2b_2 + b_1 & \hat{c}_1 &= 2 + c_1 \\ \hat{b}_0 &= b_2 + b_1 + b_0 & \hat{c}_0 &= 1 + c_1 + c_0 \end{aligned} \quad (5.66)$$

Skalierung:

Bei der Realisierung dieser Struktur wird davon ausgegangen, daß die Zustandsspeicher entsprechend den Maximalwerten nach vorne verlängert sind. Für die Weiterverarbeitung der Zustandsvariablen wird jedoch vor der nächsten Multiplikation eine Verkürzung der Zustandsspeicher notwendig, die als mathematisches Runden nach Abschnitt 4.2.2.3 ausgeführt wird.

$${}_9F_1(z) = (z^2 - 2z + 1)/C(z) \quad (5.67)$$

$${}_9F_2(z) = (z-1)/C(z)$$

$${}_9F_3(z) = 1/C(z) \quad \text{mit } \mu=v=0$$

$$b_3 = \left[\max\{ \|{}_9F_1\|_\infty, 1 \} \right]^{-1} \quad (5.68)$$

$$b_2' = b_2/b_3, \quad b_4 = 1$$

$$\hat{b}_i' = \hat{b}_i/b_3 \quad i=0,1,2 \quad (5.69)$$

Die notwendige Verlängerung der Zustandsspeicher und die damit verbundene Verschiebung nach den Integrierern berechnet sich mit Gl. (5.67) zu

$$\mu = \left\langle 1d(b_3 \cdot \|{}_9F_2\|_\infty) \right\rangle, \quad \mu \in \mathbf{Z}^{1)} \quad (5.70a)$$

$$v = \left\langle 1d(b_3^2 \cdot \|{}_9F_2\|_\infty \cdot \|{}_9F_3\|_\infty \cdot 2^{-\mu}) \right\rangle - \mu, \quad v \in \mathbf{Z} \quad (5.70b)$$

Fehlerübertragungsfunktionen:

$${}_9G_1(z) = H(z)/b_3$$

$${}_9G_2(z) = 2^\mu \frac{(\hat{b}_1' - \hat{b}_2' \hat{c}_1) z^2 + (\hat{b}_0' - 2\hat{b}_1' + \hat{b}_2' (2\hat{c}_1 - \hat{c}_0)) z + \hat{b}_1' - \hat{b}_0' + \hat{b}_2' (c_0 - c_1)}{C(z)}$$

$${}_9G_3(z) = 2^{\mu+v} \frac{(\hat{b}_0' - \hat{b}_2' \hat{c}_0) z^2 + (\hat{b}_0' (\hat{c}_1 - 2) + \hat{c}_0 (2\hat{b}_2' - \hat{b}_1')) z + (\hat{b}_0' (1 - \hat{c}_1) + \hat{c}_0 (\hat{b}_1' - \hat{b}_2'))}{C(z)}$$

$${}_9G_3(z) = 1$$

Pol-Nullstellen-Test: Bild 5.47

Abtastintervall-Test: Bild 5.48

1) $\langle a \rangle$: derart, daß gilt $0 \leq \langle a \rangle - a < 1$

b) Struktur_10 - kanonisch in Bezug auf Speicher

Signalflußgraph: Tabelle 5.3

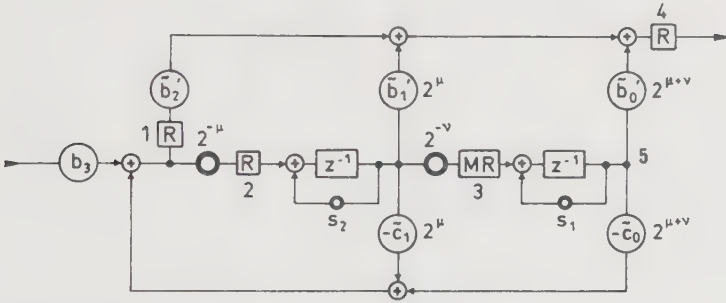


Bild 5.45: Blockschaltbild - Struktur 10

Übertragungsfunktion:

$$H(z) = \frac{\tilde{b}_2 z^2 + (\tilde{b}_1 - (s_1 + s_2) \tilde{b}_2) z + (\tilde{b}_0 - s_1 \tilde{b}_1 + s_1 s_2 \tilde{b}_2)}{z^2 + (\tilde{c}_1 - s_1 - s_2) z + (\tilde{c}_0 - s_1 \tilde{c}_1 + s_1 s_2)} = \frac{B(z)}{C(z)} \quad (5.72)$$

Koeffizienten unskaliert:

$$\begin{aligned} \tilde{b}_2 &= b_2 \\ \tilde{b}_1 &= b_1 + (s_1 + s_2) b_2 \quad ; \quad \tilde{c}_1 = c_1 + (s_1 + s_2) \\ \tilde{b}_0 &= b_0 + s_1 b_1 + s_1^2 b_2 \quad ; \quad \tilde{c}_0 = c_0 + s_1 c_1 + s_1^2 \end{aligned} \quad (5.73)$$

Skalierung:

$$\begin{aligned} {}_{10}F_1(z) &= ((z^2 - (s_1 + s_2) z + s_1 s_2) / C(z)) \\ {}_{10}F_3(z) &= (z - s_1) / C(z) \\ {}_{10}F_5(z) &= 1 / C(z) \quad \text{mit } \mu = \nu = 0 \end{aligned} \quad (5.74)$$

$$b_3 = \left[\max\{ \| {}_{10}F_1 \|_\infty, 1 \} \right]^{-1} \quad (5.75)$$

$$\begin{aligned} b_2' &= b_2 / b_3 \quad ; \quad b_4 = 1 \\ \tilde{b}_i' &= \tilde{b}_i / b_3 \quad i=0, 1, 2 \end{aligned} \quad (5.76)$$

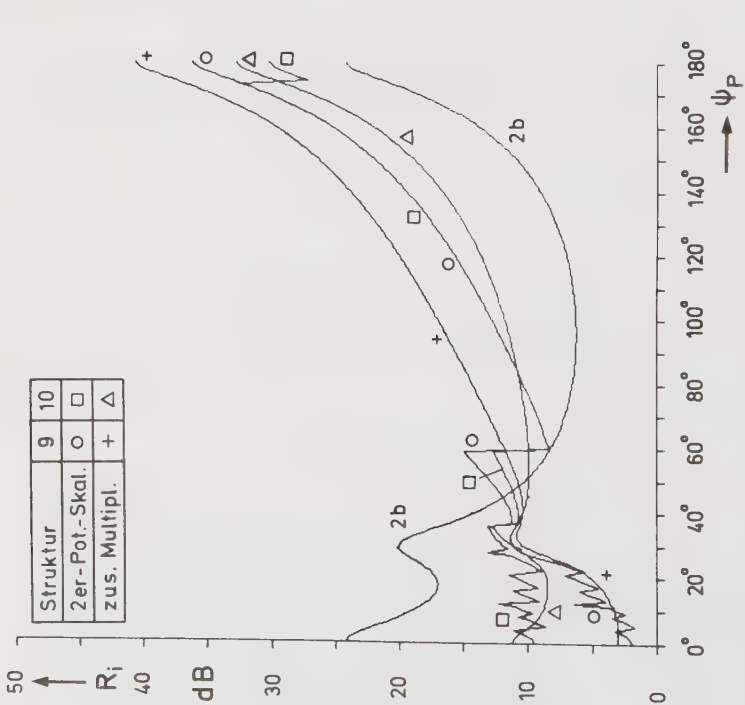


Bild 5.46: Pol-Nullstellen-Test $\psi_N = \pm 30^\circ$ $\rho_P = 0.9$
Struktur 2,9,10 ($s_1=s_2=1$)

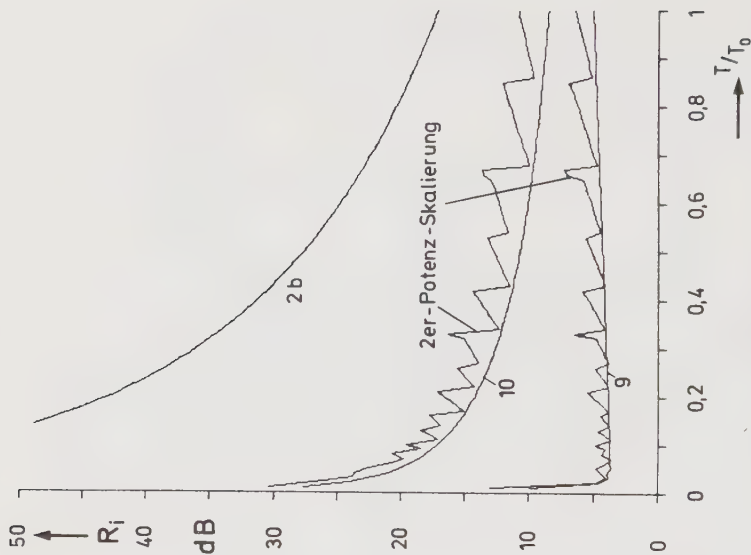


Bild 5.47: Abtastintervall-Test
Struktur 2,9,10 ($s_1=s_2=1$)

$$\begin{aligned} \mu &= \left\langle (1d \ b_3 \ || \ 10^{F_3} \ || \ \infty) \right\rangle & \mu \in \mathbb{Z} \\ \nu &= \left\langle (1d \ b_3^2 \ || \ 10^{F_3} \ || \ \infty \ || \ 10^{F_5} \ || \ \infty \cdot 2^{-\mu}) \right\rangle - \mu & \nu \in \mathbb{Z} \end{aligned} \quad (5.77)$$

Fehlerübertragungsfunktionen:

$$\begin{aligned} 10^{G_1}(z) &= \tilde{b}_2' \\ 10^{G_2}(z) &= 2^\mu \frac{(\tilde{b}_1' - \tilde{b}_2' \tilde{c}_1) z + (\tilde{b}_0' - \tilde{b}_2' \tilde{c}_0 - s_1 (\tilde{b}_1' - \tilde{b}_2' \tilde{c}_1))}{C(z)} \\ 10^{G_3}(z) &= 2^{\mu+\nu} \frac{(\tilde{b}_0' - \tilde{b}_2' \tilde{c}_0') z + (\tilde{b}_0 \tilde{c}_1 - \tilde{b}_1' \tilde{c}_0 - s_2 (\tilde{b}_0' - \tilde{b}_2' \tilde{c}_0'))}{C(z)} \\ 10^{G_4}(z) &= 1 \end{aligned} \quad (5.78)$$

Pol-Nullstellen-Test:	Bild 5.46	}	$s_1=1$
Abtast-Intervall-Test:	Bild 5.47		$s_2=1$

In den Bildern 5.46 und 5.47 werden die Ergebnisse der beiden Tests für die Strukturen 9 und 10 ($s_1=1$, $s_2=1$) aufgezeigt. Durch die innere Skalierung in Potenzen von zwei entsteht ein sprunghafter Verlauf der Rauschzahlen in Abhängigkeit vom Winkel Ψ_P . Zum Vergleich sind die Kurven der Rauschzahlen eingezeichnet, die sich ergeben, wenn anstatt der "Shift"-Multiplizierer ¹⁾echte Multiplizierer verwendet werden. Der Vergleich mit der Struktur 2b zeigt einen deutlichen Vorteil der neuen Strukturen bei kleinen Abtastintervallen bzw. Polen und Nullstellen in der Gegend von $z = 1$.

Bei freier Wahl der Parameter s_1 und s_2 sind im Bild 5.48 die jeweils günstigsten Rauschzahlen der Struktur 10 für den Pol-Nullstellen-Test bei unterschiedlichen Polradien aufgetragen. Der Vergleich mit den Kurven der 2. kanonischen Form zeigt, daß durch die Optimierung das Anwachsen der Rauschleistung für Pole in der Gegend von $z = \pm 1$ vermieden werden kann.

Einen Überblick über die günstigste Wahl der Parameter s_1 und s_2 in Abhängigkeit von den Nennerkoeffizienten c_0 , c_1 gibt das

¹⁾ "Shift"-Multiplizierer $\hat{=}$ Multiplikation mit Potenzen von zwei durch einfache Schiebeoperation

Bild 5.49. Entsprechend der untenstehenden Tabelle sind s_1 und s_2 den Kennziffern zuzuordnen.

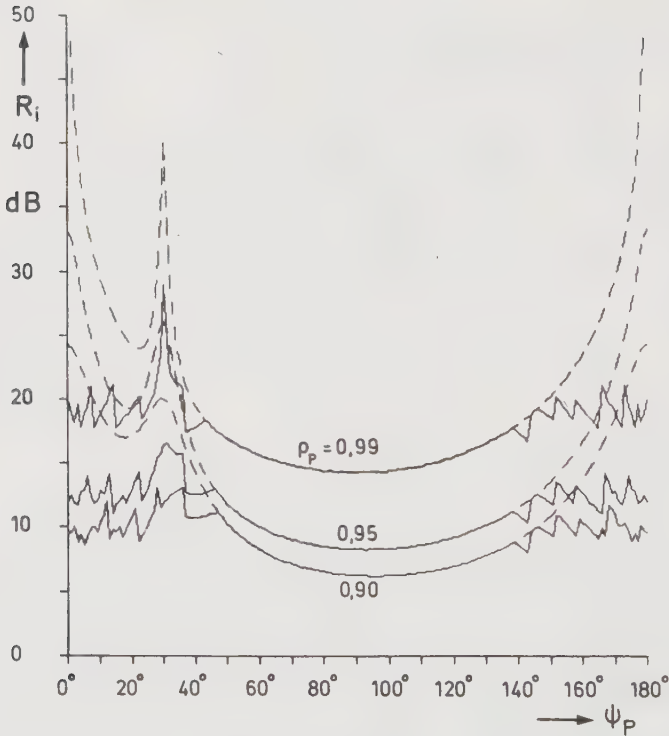


Bild 5.48: Pol-Nullstellen-Test

Struktur 10 - s_1, s_2 optimiert ———
Zum Vergleich Struktur 2 b - - - - -

5.2.7 Das Rauschverhalten bei 2er-Potenz-Skalierung mit anschließender Verkürzung

Die Skalierung der einzelnen Teilblöcke innerhalb einer Kaskaden- oder Parallelschaltung von Blöcken 2. Grades wird häufig durch "Shift"-Multiplikationen vorgenommen. Neben der generellen Verschlechterung des Rauschsignalverhältnisses um

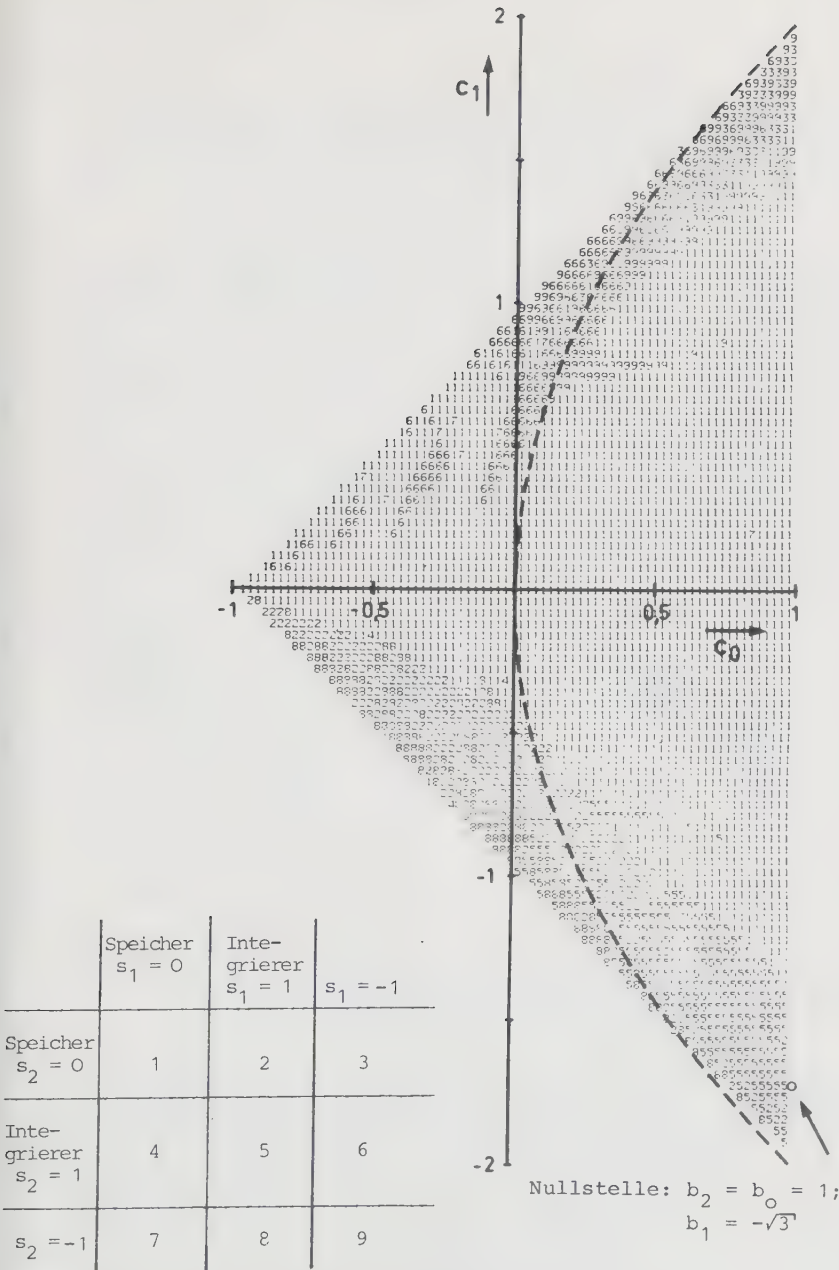


Bild 5.49: Optimale Wahl der Parameter s_1 und s_2

maximal 6dB aufgrund der ungünstigeren Aussteuerung der Teilsysteme, wirken sich hier die diskreten Fehlerverteilungen, wie sie in Abschnitt 4.2.2 beschrieben wurden, ungünstig auf das Rauschverhalten des Systems aus.

Am Beispiel der ersten kanonischen Form wird die Auswirkung einer Multiplikation mit "kurzen" Koeffizienten und anschließender Verkürzung mit unterschiedlicher Arithmetik näher untersucht. Zu einer angemessenen Bewertung des mit dem Signal korrelierten Fehleranteils wird das Eingangsspektrum der Nichtlinearität durch einen Nennerblock 2. Grades mit gleicher Pollage vorgefärbt. Der Einfluß weiterer Nichtlinearitäten innerhalb der Schaltung bleibt bei den hier durchgeführten Untersuchungen unberücksichtigt. Bild 5.50 zeigt die Schaltungsanordnung.

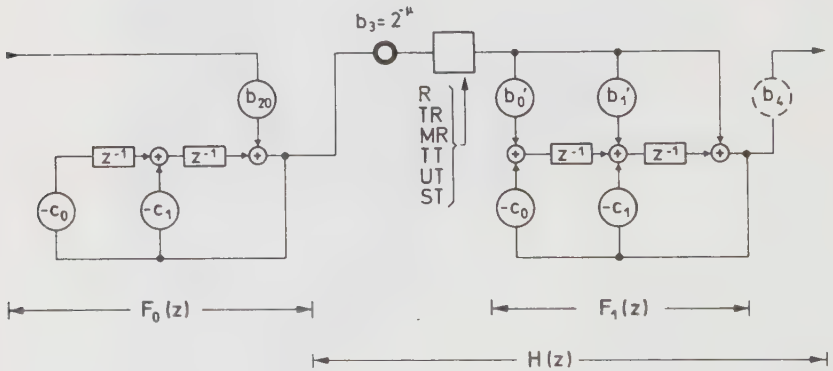


Bild 5.50: Schaltung zur Prüfung des Rauschens unterschiedlicher Skalierungsmethoden

Der Faktor b_{20} wird so gewählt, daß stets die Norm $\|F_0\|_\infty = 1$ ist. Die Übertragungsfunktion $H(z)$ wird nach Gl. 5.2 mit $\|H\|_\infty = 1$ gewählt. Der Vorfaktor $b_3 = 2^{-\mu}$ wird mit

$$\mu = \left\langle \lg \|F_1\|_\infty \right\rangle \quad (5.79)$$

und

$$F_1 = \frac{z^2 + b'_1 z + b'_0}{z^2 + c_1 z + c_0} ;$$

$$b'_0 = b_0/b_2 ,$$

$$b'_1 = b_1/b_2$$

bestimmt. Zum Ausgleich der unvollständigen Aussteuerung wird

$$b_4 = b_2/b_3 = b_2 \cdot 2^{\mu} \quad (5.80)$$

gewählt. Die Rauschzahlen der verschiedenen Nichtlinearitäten nach Abschnitt 4.2.2 werden anhand des Pol-Nullstellen-Tests miteinander und mit den Rauschzahlen eines vollständigen Skalierungsmultiplizierers und anschließender Rundung verglichen. Die logarithmische Darstellung des Faktors b_4 zeigt getrennt den Einfluß der Unteraussteuerung.

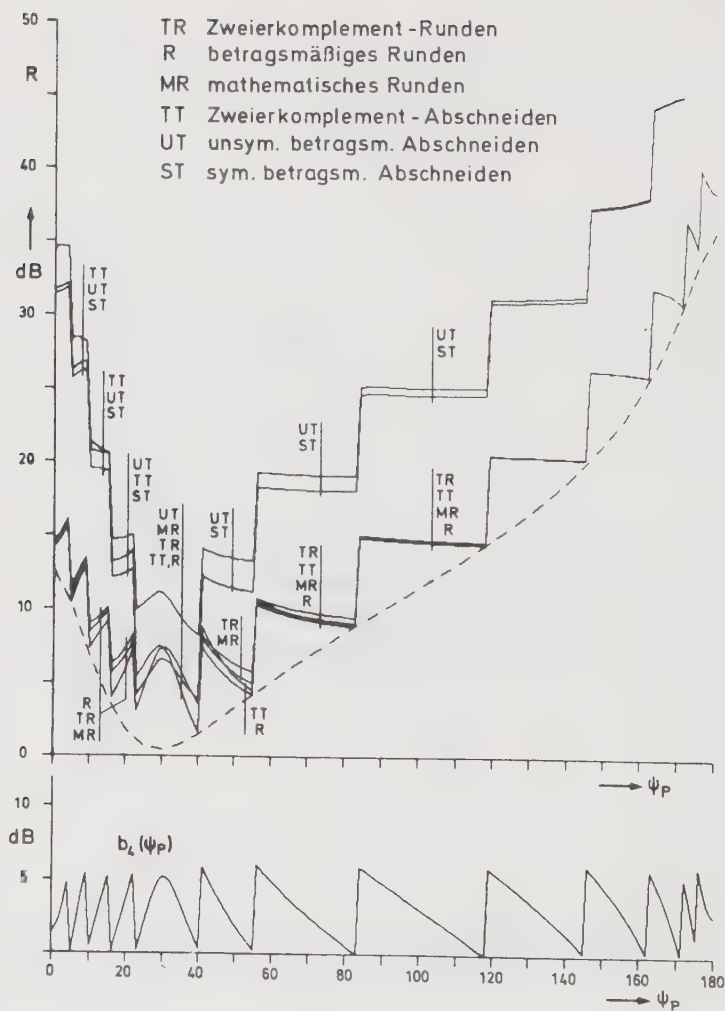


Bild 5.51: Pol-Nullstellen-Test für verschiedene Verkürzungen nach einer "Shift"-Multiplikation
Pol-Nullstellen-Test

- a) — verschiedene Verkürzungen nach einer "Shift"-Multiplikation
----- Runden nach vollst. Multiplikation
b) Einfluß der Unteraussteuerung bei "Shift"-Multiplikation.

5.3 Untersuchung und Optimierung der Kaskadenschaltung

Die Realisierung eines höhergradigen Filters kann, wie bereits einleitend bemerkt, in sehr unterschiedlicher Form erfolgen. Da ein allgemeiner Vergleich des Realisierungsaufwandes wegen der unterschiedlichen Bedingungen (z.B. Modularität, Ablaufsteuerung, Zwischenspeicher, Multiplexgrad) im Rahmen der Arbeit nicht möglich ist, bleiben die Untersuchungen hier auf die Kaskade von Blöcken 2. Grades unterschiedlicher Grundstrukturen beschränkt.

Neben den Rauscheigenschaften der verwendeten Grundstrukturen hat, wie aus der Literatur bekannt [1], [6], [71], [90] - [96] die Zuordnung der Pol- und Nullstellen der Übertragungsfunktion zu Teilblöcken sowie deren Reihenfolge einen Einfluß auf das Rauschverhalten der Gesamtanordnung. Einen Beitrag zum Auffinden der optimalen bzw. einer dieser möglichst nahe kommenden Anordnung leisten die folgenden Ausführungen. Dabei erscheint es notwendig, an einigen Beispielen die Menge aller möglichen Anordnungen auf ihr Rauschverhalten und den damit verbundenen Realisierungsaufwand hin zu prüfen, um mit Hilfe dieser Untersuchungen die verschiedenen, aus der Literatur bekannten Optimierungsverfahren überprüfen zu können und Erkenntnisse für deren Weiterentwicklung zu bekommen.

5.3.1 Verteilung und Verteilungsdichte der Aufwandsmaße bei vollständiger Enumeration aller möglichen Anordnungen der Zähler- und Nennerblöcke

Zur Bewertung der unterschiedlichen Zuordnungen von Pol-Nullstellen und deren Reihenfolge in der Kaskadenschaltung werden die inneren Rauschzahlen R_i des Gesamtsystems entsprechend Abschnitt 4.4 unter Berücksichtigung der Aussteuerung (Abschnitt 4.3) bestimmt. Dabei bleiben die Untersuchungen hier auf eine Aussteuerung entsprechend der L_∞ -Norm nach Gl. (4.91) beschränkt [97]. Die Anwendung des absoluten Übersteuerungskriteriums nach

Gl. (4.83) hat ähnliche Ergebnisse erbracht.

Die Gesamtübertragungsfunktion $H(z)$ eines höhergradigen Systems kann in $\ell = \left\lceil \frac{n+1}{2} \right\rceil$ Teilblöcke 2. Grades aufgespalten werden (Bild 5.52).

$$H(z) = b_{m_{\lambda=1}} \prod_{\lambda=1}^{\ell} \frac{z^2 + b_{1\lambda} z + b_{0\lambda}}{z^2 + c_{1\lambda} z + c_{0\lambda}} = b_{20}' \prod_{\lambda=1}^{\ell} \frac{b_{2\lambda}' z^2 + b_{1\lambda}' z + b_{0\lambda}'}{z^2 + c_{1\lambda} z + c_{0\lambda}} b_{2,\ell+1}' \quad (5.81)$$

Dabei sind die einzelnen Zählerverstärkungsfaktoren $b_{2\lambda}'$ so zu wählen, daß die Übersteuerungsgrenzen eingehalten werden.

$$b_m = \prod_{\lambda=0}^{\ell+1} b_{2\lambda}' \quad (5.82)$$

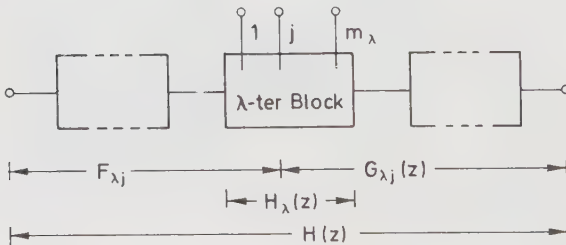


Bild 5.52: Kaskadenschaltung

Verwendet man für den Aufbau der Kaskadenschaltung Teilblöcke der ersten Art, so kann der Vorfaktor b_{20}' entfallen. Bei den Grundstrukturen der zweiten Art wird stets nur der Vorfaktor $b_{20}' = b_3$ des ersten Blocks realisiert. Die Vorfaktoren der übrigen $\ell-1$ Blöcke werden in den Zählerverstärkungsfaktor des jeweils vorausgehenden Blocks eingerechnet. Ein abschließender Faktor $b_{2,\ell+1}' > 1$ wird benötigt, wenn der letzte Block der Kaskade nicht voll angesteuert werden kann.

Alle Nichtlinearitäten werden hier entsprechend der Rundungsvorschrift nach Gl. (4.12) und gleicher Quantisierung Q_s ausgeführt. Die Rauschleistung einer Rundungsstelle beträgt $\sigma_o^2 = Q_s^2 / 12$. Die von der Quantisierungsstufe unabhängige innere Rauschzahl

R_i wird entsprechend Gln. (4.61) und (4.95) bestimmt. Enthalten die ℓ Blöcke der Schaltung jeweils m_λ Rundungsstellen mit den zugehörigen Fehlerübertragungsfunktionen $G_{\lambda j}(z)$, so ergibt sich

$$R_i = \frac{N(R)}{c_0} = \frac{1}{2\pi j} \oint_{\Gamma} \sum_{\lambda=0}^{\ell+1} \sum_{j=1}^{m_\lambda} G_{\lambda j}(z) \cdot G_{\lambda j}(z^{-1}) \frac{dz}{z} \quad (5.83)$$

Speziell erhält man für die erste und zweite kanonische Form die in den Bildern 5.53 und 5.54 dargestellt sind, bei Zusammenfassung der ungerundeten Teilprodukte

$$R_i^{(1)} = \frac{1}{2\pi j} \oint_{\Gamma} \sum_{\lambda=1}^{\ell} \frac{3}{C_\lambda(z) \cdot C_\lambda(z^{-1})} \prod_{v=\pi_\lambda+1}^{\ell} H_v(z) \cdot H_v(z^{-1}) \frac{dz}{z}, \quad (5.84)$$

$$R_i^{(2)} = \frac{1}{2\pi j} \oint_{\Gamma} \left(1 + \sum_{\lambda=1}^{\ell} \prod_{v=\lambda}^{\ell} H_v(z) \cdot H_v(z^{-1}) \right) \frac{dz}{z}, \quad (5.85)$$

$$\text{mit} \quad C_\lambda(z) = z^2 + c_{1\lambda}z + c_{0\lambda}$$

$$\text{und} \quad H_v(z) = \frac{b'_{2v}z^2 + b'_{1v}z + b'_{0v}}{C_v(z)}.$$

Je nach Wahl der Anordnung der Teilblöcke variieren die Rauschzahlen $R_i^{(1)}$ und $R_i^{(2)}$ in einem großen, sich über Zehnerpotenzen erstreckenden Bereich $|1|$. Die Anordnung mit der minimalen Rauschzahl ist die angestrebte optimale Lösung der Kaskadenschaltung. Als weiteres Bewertungsmaß wird die im Abschnitt 4.4 näher erläuterte "notwendige zusätzliche Wortlänge" Δw nach Gl. (4.101) bestimmt.

Zur Einordnung der Ergebnisse verschiedener heuristischer Optimierungsmethoden werden im folgenden die Verteilungen und die Verteilungsdichten der Werte R_i und Δw aller möglichen Realisierungen für 8 Beispiele bestimmt. Ähnliche Untersuchungen wurden in [98] für nichtrekursive Filter durchgeführt.

Für einfache Pol- und Nullstellen, die zu jeweils ℓ Paare konjugiert komplexer Polstellen und ℓ Nullstellen-Paare zusammen-

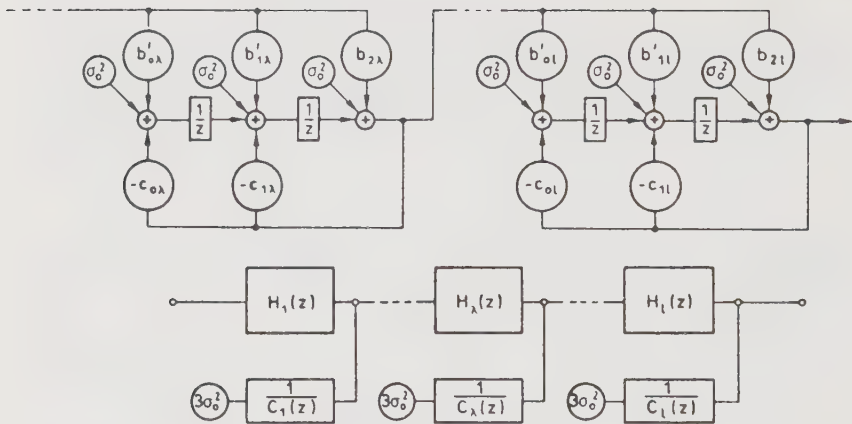


Bild 5.53: Kaskadenschaltung mit Teilblöcken
in der 1. kan. Form (Struktur 1b)

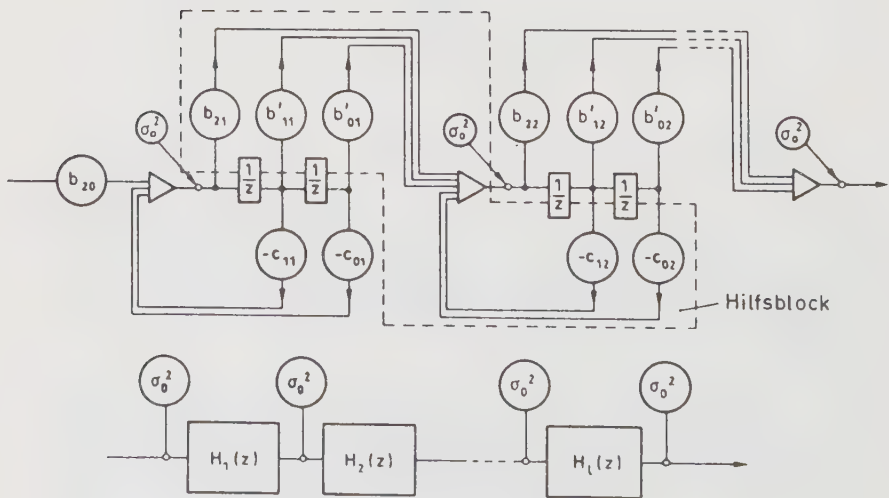


Bild 5.54: Kaskadenschaltung mit Teilblöcken
in der 2. kan. Form (Struktur 2b)

gefaßt sind, gibt es im ganzen $(\ell!)^2$ mögliche Anordnungen, die sich in der Zusammenfassung der Pol- und Nullstellen zu Teilsystemen und in der Reihenfolge dieser Teilsysteme unterscheiden. Definiert man die Verteilungsdichte, der in einen logarithmischen Maßstab (dB) umgerechneten Rauschzahlen

$$p(\rho_1) \approx \frac{\tilde{p} \{ \rho_1 \in (\rho'_1 + 0.5, \rho'_1] \}}{0.5 (\ell!)^2} \quad (5.86)$$

mit $\rho_1 = 10 \lg R_1$ und $\rho'_1 = 0.5 [2 \rho_1]$,

so erhält man einen Überblick über die Variationsbreite der Rauschzahlen bei unterschiedlicher Anordnung [97]. Im Bild 5.55a sind die Ergebnisse für die Realisierung eines Tiefpasses 10ten Grades (Anhang, Beispiel Nr. 4) in der 1. kanonischen Form aufgetragen. Das Maximum der Verteilungsdichte tritt bereits bei einem Wert $\rho_{1\max}$ auf, der nur 3 dB größer als die Rauschzahl der optimalen Anordnung $\rho_{1\text{opt}}$ ist. Bereits diese Darstellung zeigt, daß man mit relativ großer Sicherheit auch mit einfachen heuristischen Methoden eine "gute" Lösung erwarten kann. Darüber hinaus gibt die Verteilung $P_x[x \leq \Delta w]$ einen Überblick über den Anteil an Lösungen, die bei einer Verlängerung der Zustandsvariablen um Δw Bit die vorgegebene Rauschgrenze nicht überschreiten (Bild 5.55b). Wie man sieht, benötigen 7,45 % aller Lösungen (das sind 1073 von 14400) ein $\Delta w = 3$ Bit, entsprechend der notwendigen Wortlänge der optimalen Lösung. Weitere 64,7 % benötigen nur ein Bit mehr, um das Rauschen in der vorgegebenen Schranke zu halten. Entsprechende Ergebnisse für die Realisierung des Tiefpasses in der zweiten kanonischen Form zeigen die Bilder 5.56 a und b. Hier führen sogar 47,31 % aller Lösungen auf den gleichen Aufwand wie die bestmögliche Realisierung.

Ähnliche Ergebnisse konnten für eine Reihe weiterer Beispiele gefunden werden. Einen Überblick über den Abstand $\Delta \rho_{\max}$, $\Delta \rho_{\max} = \rho_{1\max} - \rho_{1\text{opt}}$ sowie den Prozentsatz an Anordnungen, die mit dem gleichen $\Delta w = \Delta w_{\text{opt}}$ oder mit einem zusätzlichen Bit erreicht werden können, gibt die Tabelle 5.7.

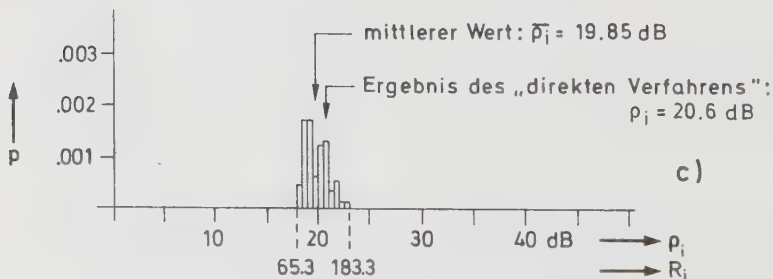
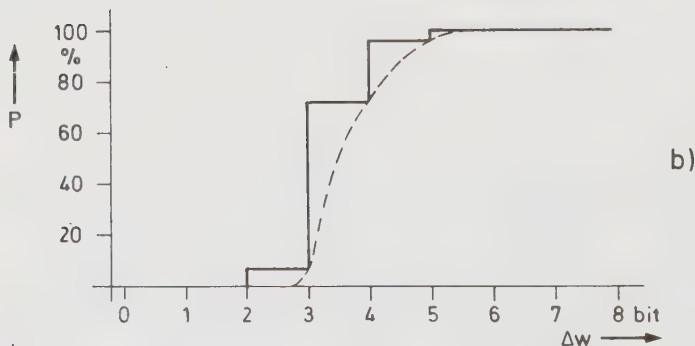
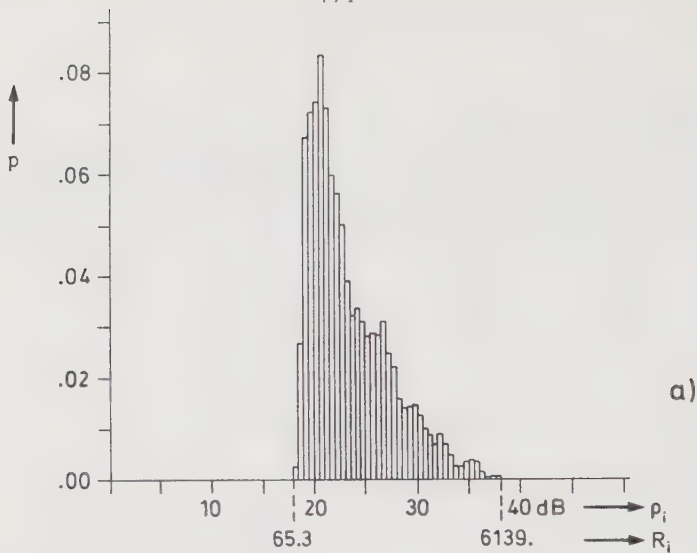


Bild 5.55: Ergebnisse der 1. kan. Form (TP 10. Grades)

- a) Verteilungsdichte $p(\rho_i)$, b) Verteilung $P_x[x \leq \Delta w]$
- c) Verteilungsdichte $p(\rho_i)$ bei fester Zuordnung

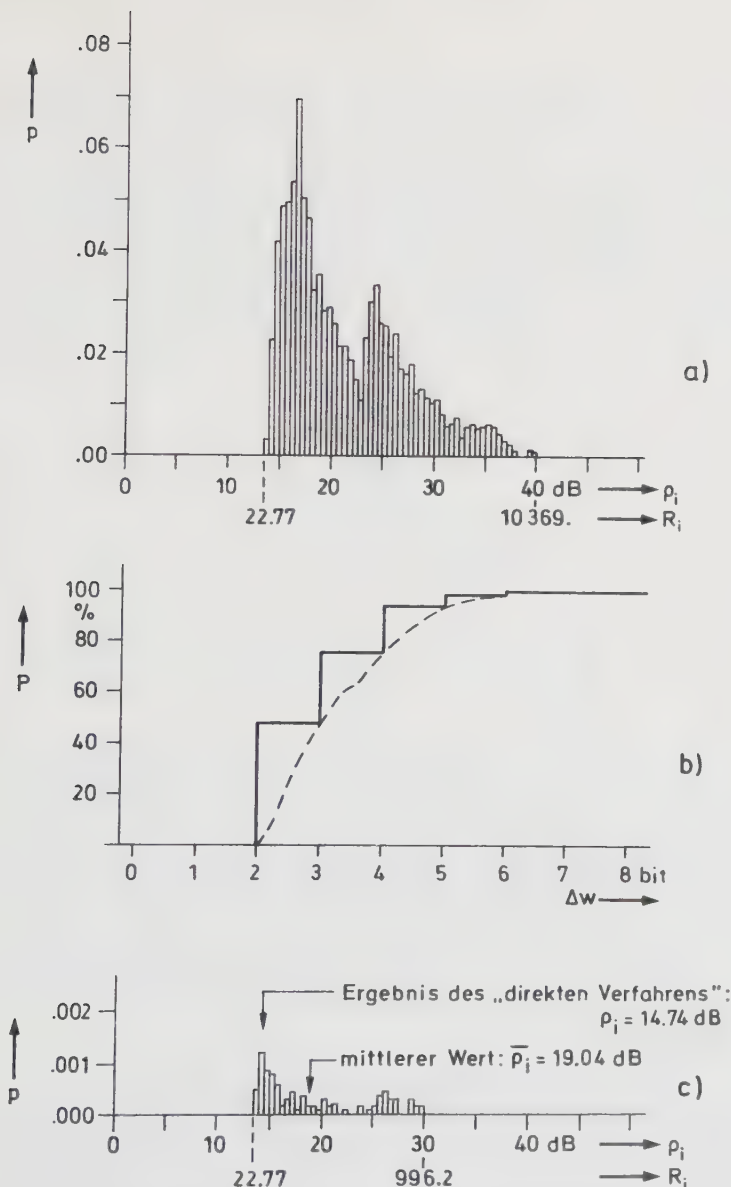


Bild 5.56: Ergebnisse der 2. kan. Form (TP 10. Grades)

- a) Verteilungsdichte $p(\rho_i)$, b) Verteilung $P_x[x \leq \Delta w]$
 c) Verteilungsdichte $p(\rho_i)$ bei fester Zuordnung

Filter	Struktur	$\Delta p_{\max}$	Anteil der Lösungen		Datensatz
			$\Delta w = \Delta w_{\text{opt}}$	$\Delta w \leq \Delta w_{\text{opt}} + 1$	
Tiefpaß 8. Grades	1b	1,5 dB	60,42 %	93,58 %	Anhang Nr. 2
	2b	2,0 dB	41,67 %	68,23 %	
Tiefpaß 10. Grades	1b	3,0 dB	7,45 %	64,7 %	Anhang Nr. 4
	2b	3,5 dB	47,31 %	76,0 %	
Bandpaß 8. Grades	1b	2,5 dB	44,62 %	79,51 %	Anhang Nr. 12
	2b	5,0 dB	18,58 %	71,70 %	
Bandpaß 10. Grades	1b	3,0 dB	33,39 %	68,76 %	Anhang Nr. 13
	2b	7,0 dB	11,6 %	57,47 %	

Tabelle 5.7: Maximum der Verteilungsdichte und Prozentsatz der Lösungen mit gleicher zusätzlicher Wortlänge Δw

Die bei der vollständigen Enumeration an Beispielen erhaltenen Ergebnisse lassen die Feststellung gerechtfertigt erscheinen, daß bereits einfache heuristische Optimierungsverfahren, die nur einen geringen Rechenaufwand und wenig Speicherplatz benötigen, zu Lösungen führen können, die in Bezug auf die notwendige zusätzliche Wortlänge Δw optimal sind. Da allein diese zusätzliche Wortlänge für den Realisierungsaufwand von Interesse ist, muß die gefundene Lösung nicht unbedingt optimal in Bezug auf die Rauschzahl sein.

In der Praxis wirkt sich dieser Quantisierungseffekt noch sehr viel stärker aus, weil in der Regel Bauteile verwendet werden müssen, bei denen die Wortlängen Vielfache von 4 oder 8 Bits sind. So wird sich auch aus dieser Sicht nur in wenigen Grenzfällen ein aufwendiges Optimierungsverfahren lohnen.

Eine erste Einschränkung der möglichen Anordnungen ergibt sich aus einer geeigneten Vorabzuordnung der Pol- und Nullstellen zu Blöcken 2. Grades, die im nächsten Abschnitt behandelt werden soll.

5.3.2 Feste Zuordnung der Pol- und Nullstellen

Vergleicht man bei der Untersuchung aller möglichen Anordnungen die verschiedenen Reihenfolgen einer festen Zuordnung, so ergeben sich für die in [1] vorgeschlagene Zuordnungsregel bei der 1. kanonischen Form besonders viele günstige Realisierungen. Die Regel wird anhand von Bild 5.57 am Beispiel des Tiefpasses 10. Grades verdeutlicht.

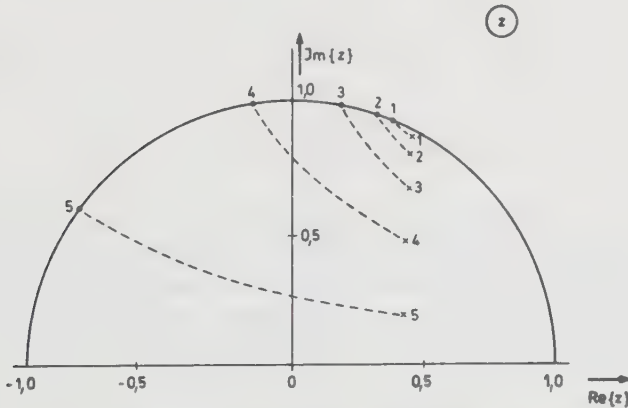


Bild 5.57: Feste Zuordnung bei den Strukturen der ersten Art und Hilfsblöcke für die Strukturen der zweiten Art

Zuordnungs-Regel:

In einem ersten Schritt werden die einander am nächsten liegenden Pol- und Nullstellen einem Block 2. Grades zugeordnet. Die weiteren Zuordnungen werden dann in gleicher Weise gewählt, ohne die bereits vergebenen Pol- und Nullstellen zu beachten.

Vergleicht man diese Zuordnung mit den Kriterien, die in [6], [90] - [92] für eine günstige Zuordnung angegeben werden, so kann man feststellen, daß die oben beschriebene Methode gerade Überhöhungen der einzelnen Übertragungsfunktionen abbaut und somit ungünstige Aussteuerungsbedingungen vermeidet.

Die Verteilungsdichte aller $\ell!$ unterschiedlichen Reihenfolgen bei entsprechend vorgegebener Zuordnung wird für den Tiefpaß 10. Grades im Bild 5.55c gezeigt. Da die optimale Lösung bei der Ausnutzung aller Freiheitsgrade der gleichen Zuordnung folgt, ist diese optimale Lösung hier in der Teilenumeration enthalten. Die ungünstigste Reihenfolge zeigt hingegen ein nur um 5 dB ungünstigeres Rauschverhalten. Eine weitere Bestätigung dieser Zuordnungsregel erhält man bei der Berechnung der Mittelwerte aller Lösungen einer bestimmten Zuordnung $\bar{\rho}_i$

$$\bar{\rho}_i = \frac{1}{\ell!} \sum_{v=1}^{\ell!} 10 \lg R_{i_v} \quad (5.87)$$

Die oben beschriebene Zuordnung liefert für das behandelte Beispiel ebenfalls den günstigsten Mittelwert.

Diese für die Strukturen der ersten Art getroffene Regel kann durch einige Überlegungen auf die Strukturen der 2. Art (z.B. 2. kanonische Form) erweitert werden [97]. Denkt man sich die Zählerfunktion eines Blockes mit der Nennerfunktion des folgenden Blockes (entsprechend der gestrichelten Linie in Bild 5.54) zu einem Hilfsblock verbunden, so kann die oben behandelte Regel auf die Hilfsblöcke angewendet werden. Zur Optimierung der Reihenfolge werden dann bei den Strukturen der 2. Gruppe die entsprechenden Hilfsblöcke vertauscht. Aufgrund der versetzten Zuordnung zur endgültigen Anordnung wird der erste Filterblock aus der Zählerfunktion des ersten Hilfsblockes und der Nennerfunktion des letzten gebildet. Die unterschiedliche Zuordnung beider Gruppen wird noch einmal durch die Schemata im Bild 5.58 verdeutlicht.

Die Verteilungsdichte der Rauschzahlen ρ_i , die mit einer derart fest vorgegebenen Zuordnung für den Tiefpaß 10. Grades in der 2. kanonischen Form gefunden wurde, ist im Bild 5.56c dargestellt. Es bleibt noch zu bemerken, daß auch hier die optimale Lösung der Regel für die feste Zuordnung folgt. Die größere Variationsbreite der Rauschzahlen bei der 2. kanonischen Form beruht auf der kritischen Wahl des ersten Nennerblocks bei dieser Struktur.

1. kanonische Form

1	2	3	4	5	Zähler
1	2	3	4	5	Nenner

2. kanonische Form

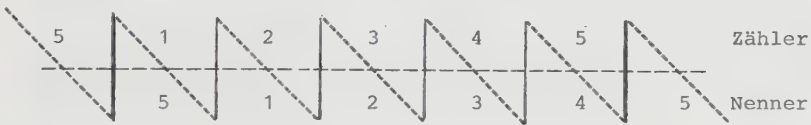


Bild 5.58: Zuordnung für Pole und Nullstellen

Die Untersuchungen aller möglichen Anordnungen bei anderen Grundstrukturen haben die Eignung dieser Grundregel auch dort bestätigt. Die Anwendung dieser Regel kann als ein erster Schritt zur Vereinfachung der Optimierungsverfahren gewertet werden.

5.3.3 Optimierung der Zuordnung und Reihenfolge der Teilblöcke

In der Literatur wurden verschiedene heuristische Optimierungsverfahren angegeben, die bereits zu recht guten Ergebnisse führen. So unterteilt Avenhaus in [6] die Optimierung in zwei Hauptteile - einen zur Bestimmung einer günstigen Zuordnung und einen anderen zur Auswahl einer geeigneten Reihenfolge. Lee [92] unterteilt ebenfalls das Optimierungsproblem in zwei Unterabschnitte nach Zuordnung und Reihenfolge. Als Optimierungskriterium wird nicht die Rauschleistung direkt, sondern das Verhältnis der maximalen zur minimalen Amplitude im Durchlaßbereich des Filters benutzt. Da heute geeignete Methoden zur Berechnung der Rauschleistung, auch bei mehrfachen Polen [83] vorliegen, erscheint ein Ausweichen auf ein "Hilfskriterium" nicht mehr erforderlich. Allgemeine Regeln für die Zuordnung und Reihenfolge

werden für die erste und zweite kanonische Form von Jackson in [71] und [90] angegeben. Die angegebenen Auswahlkriterien beziehen sich jedoch nur auf die einzelnen Blöcke 2. Grades und nicht auf das Verhalten der Blöcke im Rahmen des Gesamtfilters.

Die Forderung nach einer bei vorgegebenen Randbedingungen optimalen Lösung der Teilfilteranordnung verlangt eine gleichzeitige Behandlung der beiden Freiheitsgrade - Zuordnung und Reihenfolge - und die Betrachtung der Gesamtrauschleistung am Ausgang des Systems. Neben der überaus aufwendigen vollständigen Enumeration ermöglicht das aus der Unternehmensforschung bekannte Verfahren, das "Dynamische Programmieren" eine exakte Lösung der Aufgabe [99]. Die Anwendung dieser Methoden wurde unter anderem von Hwang [93] und Lüder et. al. [94] vorgeschlagen. Bei höhergradigen Filtersystemen ($n > 10$) steigen jedoch auch bei diesen Verfahren Rechenzeit und Speicheraufwand über alle vertretbaren Grenzen. Da im folgenden ein neues, sehr effektives heuristisches Optimierungsverfahren vorgestellt werden soll, das auf den Grundgedanken der "Dynamischen Programmierung" aufbaut, erscheint eine kurze Darstellung der exakten Methode notwendig.

5.3.3.1 Dynamisches Programmieren

Hwang führt in [93] die Akkumulation des Rauschens der einzelnen Blöcke auf einen Stufenprozess zurück. Da der Beitrag der einzelnen Stufen nur von der Stufe selbst und den vorausgehenden Optimierungsstufen abhängt, kann mit Hilfe der "Dynamischen Programmierung" eine exakte Lösung des Problems gefunden werden [100]. Bei dem hier vorliegenden Reihenfolgeproblem stellt die schrittweise Festlegung von 1,2,..., n Teilblöcken 2. Grades die n Stufen des Prozesses dar.

In einer ersten Stufe werden alle möglichen Kombinationen von jeweils einem Zähler- und Nennerblock zu einem Teilfilter 2.

Grades geprüft. Dabei wird stets die notwendige Skalierung vom Eingang bis zu der Übersteuerungsstelle durchgeführt. Bei der Berechnung der Rauschleistung am Ausgang werden die Übertragungsfunktionen der restlichen $l-1$ nachfolgenden Teilfilter berücksichtigt. Die Reihenfolge und Aussteuerung dieser restlichen Blöcke geht dabei nicht in die Berechnung ein. Die Reihenfolgen der einzelnen Anordnungen werden abgespeichert und stehen für den weiteren Verlauf der Optimierung zur Verfügung. Da im vorliegenden Fall bei der ersten Stufe alle Teilergebnisse gespeichert werden müssen und somit noch keine Selektion erfolgt, kann diese Optimierungsstufe auch mit in die nächste Stufe einbezogen werden [94].

In der zweiten Stufe werden nun alle möglichen $\binom{l}{2}^2$ Kombinationen von jeweils zwei Zähler- und zwei Nennerblöcken berücksichtigt. Für jede dieser Kombinationen ergeben sich 2^2 verschiedene Zuordnungen. Diese werden unter Berücksichtigung der $l-2$ nachfolgenden Teilblöcke überprüft und die günstigste für die weitere Optimierung abgespeichert. Bereits beim zweiten Schritt entsteht somit eine Selektion von $1:2^2$.

In der v -ten Stufe werden allgemein $\binom{l}{v}^2$ Kombinationen von jeweils v Zähler- und Nennerblöcken untersucht. Die jeweils günstigste Anordnung der ersten $v-1$ Blöcke liegt bereits aus der vorhergehenden Stufe fest und kann aus dem Speicher entnommen werden, so daß auch hier nur noch die restlichen v^2 Anordnungen des v -ten Blockes überprüft werden müssen. Es sind für die v -te Stufe somit

$$M_v = \binom{l}{v}^2 v^2 \quad (5.88)$$

Teilfilter zu untersuchen.

Das Optimierungsverfahren wird bis zur l -ten Stufe fortgesetzt. Die beste Lösung aus der letzten Stufe ist das exakte Ergebnis.

Im ganzen werden beim "Dynamischen Programmieren"

$$M = \sum_{v=2}^{\ell} (\ell_v)^2 v^2 \quad (5.89)$$

Berechnungen von Teilfiltern unterschiedlicher Länge benötigt. Gegenüber der vollständigen Enumeration bedeutet dies einen erheblichen Gewinn an Rechenzeit, wie der Vergleich der verschiedenen Verfahren im Abschnitt 5.3.3.4 zeigen wird. Bei größeren Filtersystemen steigt jedoch auch hier der Aufwand an Rechenzeit in eine, gemessen am tatsächlichen Optimierungserfolg, nicht vertretbare Größenanordnung.

Für den Ablauf der "Dynamischen Programmierung" ist nur die Abspeicherung der jeweils besten Anordnungen einer Kombination notwendig, die zur Speicherreduzierung in einer codierten Form erfolgen kann. Eine weitere Einsparung an Rechenzeit ist jedoch möglich, wenn man daneben auch das Produkt der Skalierungsfaktoren und die Teilrauschzahlen der Zwischenergebnisse abspeichert.

Bei größeren Filtersystemen steigt neben der Rechenzeit auch der Speicheraufwand rasch an und überschreitet die Kapazität eines mittleren Laborrechners. So würden z.B. für ein Filter vom Grad $n = 22$ und $\ell = 11$ bei optimaler Codierung $s > 1.28 \cdot 10^6$ Speicherplätze (Integer-Worte à 16 Bit) allein zum Zwischenspeichern der verschiedenen Anordnungen benötigt.

5.3.3.2 Optimierung mit begrenztem Zwischenspeicher

Das starke Anwachsen der Rechenzeiten und des Speicherbedarfs führt dazu, daß man bei größeren Systemen stets auf ein heuristisches Verfahren ausweichen muß. In der Literatur werden hierzu die verschiedensten Methoden [6], [92], [95] angegeben. Ein besonderer Nachteil dieser Verfahren ist darin zu sehen, daß auch bei kleineren Systemen nicht mit Sicherheit eine exakte Lösung erreicht werden kann.

Im Rahmen der aus der Unternehmensforschung bekannten "Entscheidungsbaumverfahren" [99] wird daher ein Algorithmus entwickelt,

der einen kontinuierlichen Übergang von der "Dynamischen Programmierung" zu einer "Branch and Bound"-Methode durch schrittweise Einschränkung des Zwischenspeichers ermöglicht. Bei kleineren Systemen und ausreichendem Speicherplatz wird bei rein paralleler Arbeitsweise das exakte Ergebnis erreicht. Größere Systeme arbeiten mit einem eingeschränkten Zwischenspeicher. Da nicht mehr alle Zwischenergebnisse abgespeichert werden können, geht der Algorithmus in eine sequentielle Arbeitsweise über. Das Ergebnis ist hierbei nicht mehr zwangsläufig exakt.

Das Verfahren mit begrenztem Speicherplatz baut auf der "Dynamischen Programmierung" auf. Es kann hierbei prinzipiell in gleicher Weise wie im vorausgehenden Abschnitt vorgegangen werden. Da aber nicht mehr die optimalen Anordnungen aller Kombinationen, sondern nur die $S_{\max}$ besten Lösungen abgespeichert werden, kann im nachfolgenden Schritt auch nur noch ein Teil der Kombinationen weiterverarbeitet werden. Bei der notwendigen assoziativen Speicherorganisation entsteht somit ein hoher Suchaufwand, der bei größeren Systemen in den meisten Fällen mit der Entscheidung: "Teilsystem des vorausgehenden Schritts nicht vorhanden" enden wird.

Durch eine günstigere Organisation des Ablaufs kann man erreichen, daß die Suchzeiten wesentlich verringert werden. Hierbei wird in den einzelnen Schritten nicht von den möglichen Kombinationen der Teilfilter, sondern von den vorher abgespeicherten besten Teillösungen ausgegangen.

In der ersten Stufe werden alle möglichen Zuordnungen eines Zähler- und eines Nennerblocks zu einem Teilsystem 2. Grades, gefolgt von den restlichen Teilblöcken untersucht. Die $S_{\max}$ besten Anordnungen werden zur Weiterverarbeitung bereitgestellt.

Im zweiten und den folgenden v -ten Schritten wird nacheinander von den abgespeicherten Anordnungen des vorhergehenden Schritts ausgegangen und diese um jeweils einen der $(l-v+1)$ noch nicht festgelegten Zähler- und Nennerblöcke erweitert. Nach der Berechnung der Skalierung und der Rauschleistung am Ausgang des Gesamtsystems wird überprüft, ob bereits eine Lösung dieser Kom-

bination vorhanden ist. Wenn ja, wird die beste der beiden für die weiteren Stufen abgespeichert; wenn nein, wird überprüft, ob für eine weitere Teillösung noch Platz im Speicher vorhanden ist, oder ob gegebenenfalls eine schlechtere Teillösung einer anderen Kombination überschrieben werden muß. Die Organisation des Zwischenspeichers erfolgt auch hier assoziativ.

Im 2-ten und letzten Schritt bleibt nur noch jeweils ein Zähler- und Nennerblock zum Anlagern übrig. Die Lösung mit der geringsten Rauschleistung ist das beste auf diesem Wege erreichbare Ergebnis.

Für den Ablauf des Verfahrens mit begrenztem Speicher ist bei der zuletzt beschriebenen Programmorganisation die Zwischenspeicherung der codierten Anordnungen und der Rauschleistung der zugehörigen Teilfilter notwendig. Da die Berechnung der verschiedenen Zuordnungen einer Kombination von Blöcken 2. Grades nicht direkt nacheinander erfolgt, müssen die Rauschleistungen zum Vergleich aufbewahrt werden. Eine weitere Beschleunigung der Optimierung ist auch hier durch zusätzliche Zwischenspeicherung der Produkte der Skalierungsfaktoren möglich.

Wie schon erwähnt, führt der Algorithmus bei ausreichendem Speicherplatz zur optimalen Anordnung. Begrenzt man hingegen den Zwischenspeicher auf einen einzigen Platz, $S_{\max} = 1$, so geht das Verfahren in eine endgültige (= direkte) Angliederung eines Pol-Nullstellenpaares an das Teilfilter der vorausgehenden Stufe über. Im folgenden soll dieser Grenzfall, der ein reines "Branch and Bound" darstellt, als "Direktes Verfahren" bezeichnet werden [97].

Die Anwendung und der Vergleich des Verfahrens mit begrenztem Speicherplatz zeigt, daß durch diese stark vereinfachte Optimierung gute Ergebnisse erzielt werden können. Eine weitere Reduzierung der Rechenzeit ist durch die gleichzeitige Anwendung der Zuordnungsregel möglich.

5.3.3.3 Vergleich der Ergebnisse verschiedener Verfahren

Die mit den verschiedenen Optimierungsverfahren erzielten Ergebnisse werden nun am Beispiel des im Abschnitt 3.4 vorgestellten Tiefpasses 10. Grades mit dem Übertragungsverhalten nach Bild 3.11 (Koeffizienten siehe Anhang, Beispiel Nr. 4) und an der von Golden und Kaiser [54] vorgeschlagenen Bandsperre 22. Grades (Koeffizienten siehe Anhang, Beispiel Nr. 14) verglichen.

Die Optimierungsergebnisse für den Tiefpaß 10. Grades sind in Tabelle 5.8 für die Grundstrukturen 1b (1. kanonische Form) und 2b (2. kanonische Form) dargestellt. Ein Vergleich der aufgewendeten Rechenzeiten zeigt die Effektivität der Verfahren. In Bezug auf die Rechenzeiten beim "Dynamischen Programmieren" kann der Rechenaufwand beim "direkten Verfahren" mit fester Zuordnung ungefähr auf ein hunderstel reduziert werden. Dabei werden im ersten Fall $M = 1725$, im zweiten Fall $M = 14$ Teilsysteme mit unterschiedlicher Länge berechnet. Im Fall der 2. kanonischen Form wird dabei ebenfalls die optimale Wortlänge $\Delta w = 3$ Bit erreicht.

Der Einfluß der Größe des Zwischenspeichers wird an den Ergebnissen der Tabelle 5.9 ersichtlich. Bereits mit $S_{\max} = 2$ kann auch bei der 1. kanonischen Form und fester Zuordnung die optimale Wortlänge erreicht werden. Mit $S_{\max} = 100$ geht das Verfahren bei diesem Beispiel in die "Dynamische Programmierung" über. Bei der Optimierung der Bandsperre 22. Grades kann durch eine Erweiterung des Zwischenspeichers auf $S_{\max} = 10$ eine Verbesserung der notwendigen zusätzlichen Wortlänge um 1 Bit erreicht werden. Die Angabe einer exakten Lösung ist bei diesem Beispiel aus Speicherplatzgründen nicht möglich. Auch ein direkter Vergleich mit den in [6] und [95] angegebenen Lösungen ist hier nicht möglich, da dort unterschiedliche Grundstrukturen bzw. Aussteuerungskriterien verwendet wurden. Die hier erhaltenen Ergebnisse sind in Tabelle 5.10 zusammengestellt.

	freie Zuordnung		feste Zuordnung	
	vollständige Enumeration	Dynamische Programm.	Dynamische Programm.	direktes Verfahren
S t r u k t u r Erste kanonische Form				
Anordnung	$\frac{3 \ 4 \ 1 \ 5 \ 2}{3 \ 4 \ 1 \ 5 \ 2}$	$\frac{3 \ 4 \ 1 \ 5 \ 2}{3 \ 4 \ 1 \ 5 \ 2}$	$\frac{3 \ 4 \ 1 \ 5 \ 2}{3 \ 4 \ 1 \ 5 \ 2}$	$\frac{4 \ 3 \ 5 \ 2 \ 1}{4 \ 3 \ 5 \ 2 \ 1}$
Skalierungs- faktoren	0.362151 0.262168 0.567443	0.362151 0.262168 0.567443	0.362151 0.262168 0.567443	0.231573 0.409997 0.125002
$b_{2\lambda}$ $\lambda=1(1)5$	0.183345 0.705262	0.183345 0.705262	0.183345 0.705262	0.705409 0.832115
$R_i^{(1)} =$	65,30 $\hat{=}$ 18,15dB	65,30 $\hat{=}$ 18,15dB	65,30 $\hat{=}$ 18,15dB	115,15 $\hat{=}$ 20,61dB
Δw	$\langle 2,80 \rangle = 3 \text{ Bit}$	$\langle 2,80 \rangle = 3 \text{ Bit}$	$\langle 2,80 \rangle = 3 \text{ Bit}$	$\langle 3,20 \rangle = 4 \text{ Bit}$
Rechenzeit * min : sec	$\approx 175:00$	4:22	0:10.94	0:2.32
Zweite kanonische Form				
Anordnung	$\frac{3 \ 5 \ 1 \ 2 \ 4}{4 \ 3 \ 5 \ 1 \ 2}$	$\frac{3 \ 5 \ 1 \ 2 \ 4}{4 \ 3 \ 5 \ 1 \ 2}$	$\frac{3 \ 5 \ 1 \ 2 \ 4}{4 \ 3 \ 5 \ 1 \ 2}$	$\frac{5 \ 3 \ 2 \ 1 \ 4}{4 \ 5 \ 3 \ 2 \ 1}$
Skalierungs- faktoren	0.436469 0.371125 0.167088	0.436469 0.371125 0.167088	0.435469 0.371125 0.167088	0.436469 0.128182 0.483770
$b_{2\lambda}$ $\lambda=0(1)5$	0.716357 0.462666 0.776596	0.716357 0.462666 0.776596	0.716357 0.462666 0.776596	0.623878 0.531248 0.776596
$R_i^{(2)}$	22,77 $\hat{=}$ 13,57dB	22,77 $\hat{=}$ 13,57dB	22,77 $\hat{=}$ 13,57dB	29,80 $\hat{=}$ 14,74dB
Δw	$\langle 2,04 \rangle = 3 \text{ Bit}$	$\langle 2,04 \rangle = 3 \text{ Bit}$	$\langle 2,04 \rangle = 3 \text{ Bit}$	$\langle 2,23 \rangle = 3 \text{ Bit}$
Rechenzeit * min : sec	$\approx 197:00$	4:54	0:12.8	0:2.68

* PDP 11/45

Tabelle 5.8: Ergebnisse bei versch. Optimierungsverfahren

	$S_{\max}$	Anordnung	$R_i \hat{=} \rho_i / \text{dB}$	ΔW	Rechenzeit min : sec
Zuordnung frei	1	$\begin{smallmatrix} 5 & 4 & 3 & 2 & 1 \\ 4 & 3 & 5 & 1 & 2 \end{smallmatrix}$	$76,89 \hat{=} 18,86 \text{ dB}$	$\langle 2,92 \rangle = 3$	0:9,20
	2	$\begin{smallmatrix} 5 & 4 & 2 & 3 & 1 \\ 4 & 3 & 2 & 5 & 1 \end{smallmatrix}$	$77,77 \hat{=} 18,91 \text{ dB}$	$\langle 2,96 \rangle = 3$	0:13,56
	5	$\begin{smallmatrix} 5 & 4 & 3 & 2 & 1 \\ 4 & 3 & 5 & 1 & 2 \end{smallmatrix}$	$76,89 \hat{=} 18,86 \text{ dB}$	$\langle 2,92 \rangle = 3$	0:26,78
	10	$\begin{smallmatrix} 5 & 2 & 4 & 1 & 3 \\ 4 & 2 & 5 & 1 & 3 \end{smallmatrix}$	$67,24 \hat{=} 18,28 \text{ dB}$	$\langle 2,82 \rangle = 3$	0:46,18
	20	$\begin{smallmatrix} 5 & 2 & 4 & 1 & 3 \\ 4 & 2 & 5 & 1 & 3 \end{smallmatrix}$	$67,24 \hat{=} 18,28 \text{ dB}$	$\langle 2,81 \rangle = 3$	1:36,30
	50	$\begin{smallmatrix} 3 & 4 & 1 & 5 & 2 \\ 3 & 4 & 1 & 5 & 2 \end{smallmatrix}$	$65,30 \hat{=} 18,15 \text{ dB}$	$\langle 2,80 \rangle = 3$	2:49,08
	100	$\begin{smallmatrix} 3 & 4 & 1 & 5 & 2 \\ 3 & 4 & 1 & 5 & 2 \end{smallmatrix}$	$65,30 \hat{=} 18,15 \text{ dB}$	$\langle 2,80 \rangle = 3$	4:22,54
Zuordnung fest	1	$\begin{smallmatrix} 4 & 3 & 5 & 2 & 1 \\ 4 & 3 & 5 & 2 & 1 \end{smallmatrix}$	$115,15 \hat{=} 20,61 \text{ dB}$	$\langle 3,21 \rangle = 4$	0:2,34
	2	$\begin{smallmatrix} 3 & 4 & 2 & 1 & 5 \\ 3 & 4 & 2 & 1 & 5 \end{smallmatrix}$	$80,24 \hat{=} 19,04 \text{ dB}$	$\langle 2,94 \rangle = 3$	0:3,74
	5	$\begin{smallmatrix} 3 & 4 & 5 & 1 & 2 \\ 3 & 4 & 1 & 5 & 2 \end{smallmatrix}$	$65,30 \hat{=} 18,15 \text{ dB}$	$\langle 2,80 \rangle = 3$	0:7,84
	10	$\begin{smallmatrix} 3 & 4 & 1 & 5 & 2 \\ 3 & 4 & 1 & 5 & 2 \end{smallmatrix}$	$65,30 \hat{=} 18,15 \text{ dB}$	$\langle 2,80 \rangle = 3$	0:10,86

Tabelle 5.9: Unterschiedliche Größe des Speichers; TP10 Gr. (Struktur 1b) mit freier und fester Zuordnung

$S_{\max}$	Anordnung	$R_i \hat{=} \rho_i / \text{dB}$	ΔW	Rechenzeit min : sec
1	$\begin{smallmatrix} 6 & 5 & 7 & 4 & 8 & 3 & 2 & 1 & 9 & 10 & 11 \\ 6 & 5 & 7 & 4 & 8 & 3 & 2 & 1 & 9 & 10 & 11 \end{smallmatrix}$	$2,40 \cdot 10^4 \hat{=} 43,80$	$\langle 6,81 \rangle = 7$	0:31,60
2	$\begin{smallmatrix} 5 & 7 & 4 & 8 & 3 & 9 & 2 & 6 & 10 & 11 & 1 \\ 5 & 7 & 4 & 8 & 3 & 9 & 2 & 6 & 10 & 11 & 1 \end{smallmatrix}$	$1,26 \cdot 10^4 \hat{=} 41,01$	$\langle 6,34 \rangle = 7$	0:55,82
5	$\begin{smallmatrix} 7 & 4 & 5 & 8 & 9 & 3 & 2 & 10 & 1 & 11 & 6 \\ 7 & 4 & 5 & 8 & 9 & 3 & 2 & 10 & 1 & 11 & 6 \end{smallmatrix}$	$9719,3 \hat{=} 39,88$	$\langle 6,15 \rangle = 7$	2: 8,46
10	$\begin{smallmatrix} 4 & 8 & 6 & 9 & 3 & 1 & 5 & 10 & 11 & 7 & 2 \\ 4 & 8 & 6 & 9 & 3 & 1 & 5 & 10 & 11 & 7 & 2 \end{smallmatrix}$	$7067,2 \hat{=} 38,49$	$\langle 5,93 \rangle = 6$	4: 7,88
20	$\begin{smallmatrix} 6 & 9 & 4 & 7 & 10 & 3 & 1 & 8 & 11 & 5 & 2 \\ 6 & 9 & 4 & 7 & 10 & 3 & 1 & 8 & 11 & 5 & 2 \end{smallmatrix}$	$5398,9 \hat{=} 37,32$	$\langle 5,73 \rangle = 6$	7:12,54
50	$\begin{smallmatrix} 7 & 3 & 9 & 5 & 2 & 10 & 4 & 8 & 1 & 11 & 6 \\ 7 & 3 & 9 & 5 & 2 & 10 & 4 & 8 & 1 & 11 & 6 \end{smallmatrix}$	$4779,8 \hat{=} 36,79$	$\langle 5,64 \rangle = 6$	16:12,98
100	$\begin{smallmatrix} 5 & 8 & 2 & 10 & 6 & 3 & 9 & 4 & 11 & 1 & 7 \\ 5 & 8 & 2 & 10 & 6 & 3 & 9 & 4 & 11 & 1 & 7 \end{smallmatrix}$	$4196,6 \hat{=} 36,23$	$\langle 5,55 \rangle = 6$	28:18,08

Tabelle 5.10: Unterschiedliche Größe des Speichers;
Bandsperr 22-ten Grades (nach Kaiser) (Struktur 1b)

5.3.3.4 Abschätzung der Rechenzeiten

Die wenigen, in den vorhergehenden Abschnitten berechneten Beispiele haben gezeigt, daß die unterschiedlichen Verfahren sehr verschiedene Rechenzeiten benötigen. Um den numerischen Aufwand der einzelnen Optimierungsverfahren auch bei höhergradigen Filtersystemen im voraus vergleichen zu können, werden die nötigen Rechenzeiten abgeschätzt.

Erfolgt die Aussteuerung des Systems entsprechend der L_{∞} -Norm (nach Gl. 4.91), so müssen die Maxima der einzelnen Teilübertragungsfunktionen vom Eingang des Systems zu den Übersteuerungspunkten hin berechnet werden. Um die oft steilen Spitzen der Übertragungsfunktionen schnell und mit großer Sicherheit bestimmen zu können, wird ein iteratives Verfahren angewandt, das von wenigen, durch Projektion der Polstellen auf den Einheitskreis gewonnenen Stützstellen ausgeht. Wenngleich dieses Verfahren einer festen Rasterung bei weitem überlegen ist, stellt sich dennoch heraus, daß ein großer Teil der Rechenzeit des gesamten Optimierungsvorgangs hierfür benötigt wird.

Vergleichsweise gering ist die Rechenzeit für alle anderen Rechenoperationen. Dies gilt auch für die Bestimmung der Rauschzahl selbst, solange nur unkorrelierte Rauschquellen im System enthalten sind. Erfolgt die Aussteuerung nach dem absoluten Kriterium, der Summe der Beträge der Impulsantwort (Gl. 4.83), so steigt die Rechenzeit für die Bestimmung des Skalierungsfaktors noch weiter an und die oben gemachten Annahmen sind somit ebenfalls gerechtfertigt.

Anhand der für die Bestimmung der Skalierungsfaktoren notwendigen Rechenzeit wird eine stark vereinfachte Abschätzung der Optimierungszeiten angegeben. Die unterschiedliche Organisation der Programme - z.B. unterschiedliche Suchzeiten beim assoziativen Speicheraufruf - können dabei zu Abweichungen führen, die jedoch im Rahmen der Abschätzung unberücksichtigt bleiben.

Zur weiteren Vereinfachung wird davon ausgegangen, daß die Anzahl der im Mittel notwendigen Suchschritte zur Bestimmung des Maximums der Übertragungsfunktion unabhängig vom Grad der Übertragungsfunk-

tion und von der Art des Filter ist. Wird nun mit t_0 die durchschnittliche Rechenzeit bezeichnet, die zum Auffinden des Maximums eines Teilsystems 2. Grades benötigt wird, so benötigt die Kaskade von λ Blöcken die Zeit $\lambda \cdot t_0$. Für den Fall, daß nur ein Übersteuerungspunkt je Block 2. Grades abzufragen ist (1. kanonische Form) werden die theoretischen Optimierungszeiten für die verschiedenen Verfahren bestimmt.

a) Vollständige Enumeration

Die vollständige Enumeration benötigt somit für die $(\ell!)^2$ verschiedenen Anordnungen eine Rechenzeit T_1

$$T_1 = (\ell!)^2 t_0 \sum_{\lambda=1}^{\ell} \lambda = t_0 (\ell!)^2 \frac{\ell(\ell+1)}{2} \quad (5.90)$$

b) "Dynamische Programmierung" - freie Zuordnung

Entsprechend Gl. (5.89) sind bei der "Dynamischen Programmierung"

$$M = \sum_{v=2}^{\ell} \binom{\ell}{v}^2 v^2$$

Teilfilter unterschiedlicher Länge zu berechnen. Wird bei den einzelnen Stufen nur die Codierung der jeweils besten Teillösungen abgespeichert, so muß die Skalierung der bereits festgelegten Blöcke neu bestimmt werden und man erhält eine Gesamtrechenzeit T_2

$$T_2 = t_0 \sum_{v=2}^{\ell} \binom{\ell}{v}^2 v^2 \sum_{\lambda=1}^v \lambda = t_0 \cdot \sum_{v=2}^{\ell} \binom{\ell}{v}^2 v^3 \frac{(v+1)}{2} \quad (5.91)$$

Bei ausreichender Speicherkapazität können neben den Codierungen der einzelnen Anordnungen auch die Produkte der Skalierungsfaktoren für das jeweilige Zwischenergebnis abgespeichert werden. In diesem Fall wird nur die Bestimmung des v -ten Skalierungsfaktors notwendig. Die Rechenzeit hierfür beträgt $v t_0$. Da in einem vorausgehenden Schritt die Skalierungen der ℓ^2 möglichen Zuordnungen eines Zähler-

und Nennerblocks erfolgen muß, wird die Summation hier entsprechend von $v = 1(1)l$ durchgeführt. Man erhält eine reduzierte Gesamtzeit T_2'

$$T_2' = t_0 \sum_{v=1}^l \binom{l}{v}^2 v^2 v = t_0 \sum_{v=1}^l \binom{l}{v}^2 v^3. \quad (5.92)$$

c) "Dynamische Programmierung" - feste Zuordnung

Die Anzahl der zu berechnenden Teilfilter reduziert sich bei der festen Zuordnung auf $M' = \sum_{v=2}^l \binom{l}{v} v$. Entsprechend reduzieren sich auch die Rechenzeiten T_3 und T_3' für die Optimierung der gesamten Anordnungen.

$$T_3 = t_0 \sum_{v=2}^l \binom{l}{v} v \sum_{\lambda=1}^v \lambda = t_0 \cdot \sum_{v=2}^l \binom{l}{v} \frac{v^2(v+1)}{2} \quad (5.93)$$

$$T_3' = t_0 \sum_{v=1}^l \binom{l}{v} v v = t_0 \sum_{v=1}^l \binom{l}{v} v^2 \quad (5.94)$$

d) Optimierung mit begrenztem Speicher - freie Zuordnung

Bei unbegrenzter Speicherkapazität werden bei der "Dynamischen Programmierung" im v -ten Schritt

$$M_v = \binom{l}{v}^2 v^2 = \binom{l}{v-1}^2 (l-v+1)^2$$

Teilfilter, ausgehend von den $\binom{l}{v-1}^2 \leq S_{\max}$ besten Lösungen des vorausgehenden Schritts berechnet. Bei einer auf $S_{\max}$ begrenzten Speicherkapazität stehen für die Weiterverarbeitung jedoch nur maximal $S_{\max}$ vorausgehende Ergebnisse zur Verfügung und man erhält somit für die Gesamtrechenzeit T_4

$$\begin{aligned} T_4 &= t_0 \sum_{v=1}^l \text{MIN} \left\{ \binom{l}{v-1}^2, S_{\max} \right\} \cdot (l-v+1)^2 \cdot \sum_{\lambda=1}^v \lambda \\ &= t_0 \sum_{v=1}^l \text{MIN} \left\{ \binom{l}{v-1}^2, S_{\max} \right\} \frac{(l-v+1)^2 v(v+1)}{2}. \end{aligned} \quad (5.95)$$

Die Summation über $v = 1(1)l$ ist auch hier notwendig, da bei $S_{\max} < l^2$ bereits im ersten Schritt eine Auswahl erfolgen kann. Entsprechend erhält man bei Zwischenspeicherung der Produkte der Skalierungsfaktoren

$$T_4' = t_0 \sum_{v=1}^{\ell} \text{MIN} \{ (\binom{\ell}{v-1})^2, S_{\max} \} (\ell-v+1)^2 v \quad (5.96)$$

e) Optimierung mit begrenztem Speicher - feste Zuordnung

Entsprechende Rechenzeiten ergeben sich bei fester Zuordnung der Pol- und Nullstellen. Man erhält

$$T_5 = t_0 \sum_{v=1}^{\ell} \text{MIN} \{ (\binom{\ell}{v-1}), S_{\max} \} \frac{(\ell-v+1)v(v+1)}{2} \quad (5.97)$$

$$T_5' = t_0 \sum_{v=1}^{\ell} \text{MIN} \{ (\binom{\ell}{v-1}), S_{\max} \} (\ell-v+1) v \quad (5.98)$$

f) Direktes Verfahren

Das "Direkte Verfahren" kann sowohl bei fester als auch bei freier Zuordnung der Pol- und Nullstellen verwendet werden. Es stellt, wie oben beschrieben, einen Grenzfall der Optimierung mit begrenztem Speicher für $S_{\max} = 1$ dar. Besonders gute Optimierungsergebnisse konnten in Verbindung mit der festen Zuordnung der Pol- und Nullstellen erzielt werden. Die Rechenzeiten für diesen speziellen Fall ergeben sich aus den Gl. (5.97) und (5.98). Dabei ist zu berücksichtigen, daß nach dem $(\ell-1)$ -ten Schritt bereits das Endergebnis festliegt. Die Summation erfolgt daher ebenfalls nur über $(\ell-1)$ Stufen. Die Abschätzung der Rechenzeit T_6 für das "Direkte Verfahren" ohne Zwischenspeicherung wird dabei nur der Vollständigkeit halber angeführt, da der Rechenzeitgewinn bei Zwischenspeicherung in keinem Vergleich zur Einsparung der Speicherplätze steht.

$$T_6 = t_0 \sum_{v=1}^{\ell-1} (\ell-v+1) \frac{v(v+1)}{2} = t_0 \frac{\ell(\ell^2-1)(\ell+6)}{24} \quad (5.99)$$

$$T_6' = t_0 \sum_{v=1}^{\ell-1} (\ell-v+1) v = t_0 \frac{(\ell-1)\ell(\ell+4)}{6} \quad (5.100)$$

Zum Vergleich der unterschiedlichen Rechenzeiten wurde in Bild 5.59 die geschätzte Rechenzeit über den Filtergrad des zu optimierenden Systems aufgetragen. Berücksichtigt man den logarithmischen Maßstab der Abbildung, so wird der erhebliche Rechenzeitgewinn deutlich, z.B. verringert sich die Rechenzeit bei einem Filter 16-ten Grades bei Anwendung des "Direkten Verfahrens" gegenüber dem "Dynamischen Programmieren" mit freier Zuordnung um den Faktor 10^{-4} .

Die Rechenzeiten bei begrenztem Speicher liegen bei fester Zuordnung zwischen den Kurven T_3 und T_6 bzw. T'_3 und T'_6 .

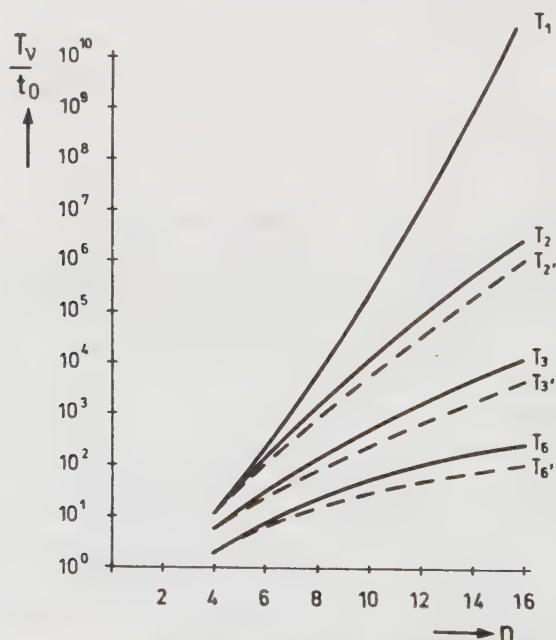


Bild 5.59: Rechenzeit bei verschiedenen Optimierungsverfahren

T_1	vollständige Enumeration
T_2, T'_2	Dynamische Programmierung Zuord. frei
T_3, T'_3	Dynamische Programmierung Zuord. fest
T_6, T'_6	Direktes Verfahren Zuord. fest
T_v ohne, T_v mit Abspeichern der Produkte der Vorfaktoren.	

5.3.3.5 Einschränkungen bei der Optimierung von Strukturen mit korrelierten Fehleranteilen

Verwendet man an mehreren Stellen innerhalb einer Struktur betragsmäßiges Abschneiden, so treten, wie bereits im Abschnitt 4.2.3 beschrieben, Korrelationen der verschiedenen vom Signal abhängigen Fehlerquellen auf. Diese Korrelationen sind nicht auf den einzelnen Block 2. Grades beschränkt. Dadurch wird eine der wesentlichsten Voraussetzungen - das Bewertungsmaß darf nur von der jeweiligen zu optimierenden Stufe oder von den vorausgehenden Stufen abhängen - für die Anwendung eines Entscheidungsbaumverfahrens verletzt. Im Rahmen der aus der Unternehmensforschung bekannten Optimierungsverfahren ist das vorliegende Problem in die Klasse der "Raumzuordnungsprobleme" einzureihen [99]. Eine exakte Lösung ist nur durch eine vollständige Enumeration möglich. Darüberhinaus sind eine Vielzahl suboptimaler heuristischer Lösungswege bekannt.

Eine Untersuchung der Verteilungen und Verteilungsdichten der Bewertungsmaße beim Aufsuchen aller möglichen Anordnungen hat gezeigt, daß auch hier die bei "Runden" gemachten Beobachtungen bestätigt werden können:

- relativ große Verteilungsdichte der möglichen Anordnungen im Bereich der der optimalen Lösung entsprechenden notwendigen zusätzlichen Wortlänge Δw ;
- günstige Zuordnung entsprechend der Regel nach Abschnitt 5.3.2.

Als Beispiel werden die Verteilung und Verteilungsdichte

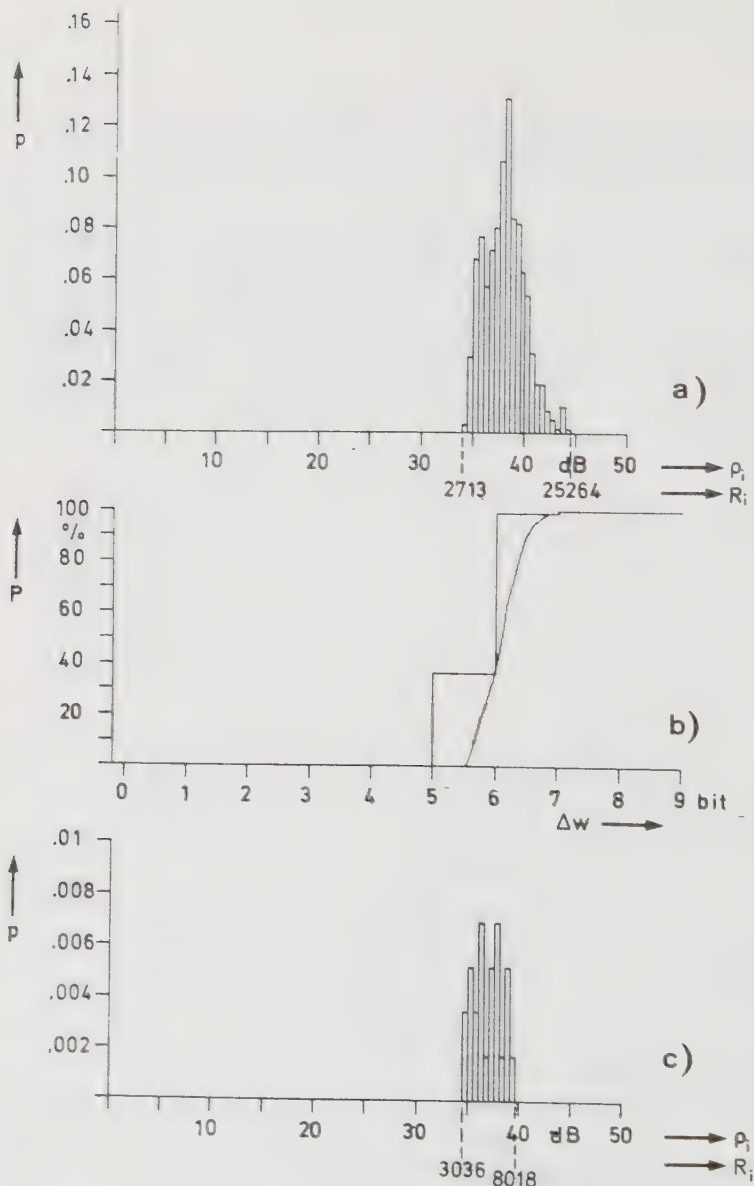


Bild 5.60: Ergebnisse der Struktur 11 (TP 8. Grades)

- a) Verteilungsdichte $p(\rho_i)$, b) Verteilung $P_x | x \leq \Delta w$
 c) Verteilungsdichte $p(\rho_i)$ bei fester Zuordnung.

der Bewertungsmaße des Tiefpasses 8. Grades bei der Grundstruktur 11 (Meerkötter, Wegener) in Bild 5.60 gezeigt. Die beste Lösung bei fester Zuordnung liegt hier um 0.48 dB schlechter als die optimale Lösung.

Bei der Überprüfung des Filters mit Blöcken 2. Grades nach Struktur 1c entsprach die optimale Lösung der Zuordnungsregel.

		Block 1			Block 2			Block 3			Block 4		
		T ₁	T ₂	R	T ₁	T ₂	R	T ₁	T ₂	R	T ₁	T ₂	R
Block 1	T ₁	4,85	5,94		1,63	2,27		0,64	-0,24		2,07	4,32	
	T ₂	5,94	14,04		3,21	6,87		1,89	3,74		1,26	2,10	
	R			2,98									
Block 2	T ₁	1,63	3,21		1,99	3,17		0,97	0,97		-0,18	-0,43	
	T ₂	2,27	6,87		3,17	11,58		2,61	7,17		-1,72	-4,46	
	R						2,83						
Block 3	T ₁	0,64	1,89		0,97	2,61		3,72	4,10		-0,82	-1,39	
	T ₂	-0,24	3,74		0,97	7,17		4,10	25,49		-3,34	-7,34	
	R									8,89			
Block 4	T ₁	2,07	1,26		-0,18	-1,72		-0,82	-3,34		10,41	16,71	
	T ₂	4,32	2,10		-0,43	-4,46		-1,39	-7,34		16,71	38,82	
	R												6,45

21.8
 $R_i^*/12 = 191.9$

Tabelle 5.11: Rauschleistungsanteile der Auto- und Kreuzkorrelationen (Struktur 11) in Q^2

Zur Entwicklung eines geeigneten heuristischen Optimierungsverfahrens werden weiter die Leistungsanteile der einzelnen Korrelationen für die optimale Lösung des Tiefpasses 8. Grades (Struktur 11, Bild 5.40) betrachtet. In Tab. 5.11 sieht man, daß neben den Autokorrelationsteilen $N_i^{(R)}$ nach Gl. (4.61) und $N_{ii}^{(S)}$ nach Gl. (4.66), die auf der Diagonalen (schraffiert) dargestellt sind, die größten Kreuzkorrelationsanteile $N_{ij}^{(S)}$ innerhalb der Blöcke 2. Grades auftreten (stark umrahmt).

Berücksichtigt man bei der Optimierung der Anordnungen ausschließlich die Autokorrelationen und die Kreuzkorrelationen innerhalb der Blöcke 2. Grades (Rauschzahl R_i'), so ist die Anwendung einer der in Abschnitt 5.3.3 beschriebenen Verfahren auch hier möglich. Wenngleich die optimale Anordnung nicht mit Sicherheit gefunden werden kann, so zeigen die Ergebnisse in Tabelle 5.12, daß auch hier mit einfachen heuristischen Verfahren sehr günstige Anordnungen erreicht werden können.

	freie Zuordnung		feste Zuordnung	
	vollständige Enumeration	dynamische Programmierung	dynamische Programmierung	direktes Verfahren
Anordnung	$\begin{array}{cccc} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{array}$	$\begin{array}{cccc} 2 & 3 & 4 & 1 \\ 2 & 3 & 4 & 1 \end{array}$	$\begin{array}{cccc} 2 & 3 & 4 & 1 \\ 2 & 3 & 4 & 1 \end{array}$	$\begin{array}{cccc} 3 & 4 & 2 & 1 \\ 3 & 4 & 2 & 1 \end{array}$
Skalierungsfaktoren $b_{2\lambda}$	0.364706 0.429537 0.174090	0.259553 0.279168 0.135202	0.259553 0.279168 0.135202	0.187651 0.088005 0.593220
$\lambda = 1(1)4$	0.281581	0.783869	0.783869	0.783869
R_i'		1961,63	1961,63	2332,16
$R_i \hat{=} \rho_i$	2713,84 $\hat{=} 34,34$ dB	3036,13 $= 34,83$ dB	3036,13 $= 34,82$ dB	3506,65 $= 35,45$ dB
$\Delta w/\text{Bit}$	<5,51> = 6	<5,58> = 6	<5,58> = 6	<5,68> = 6

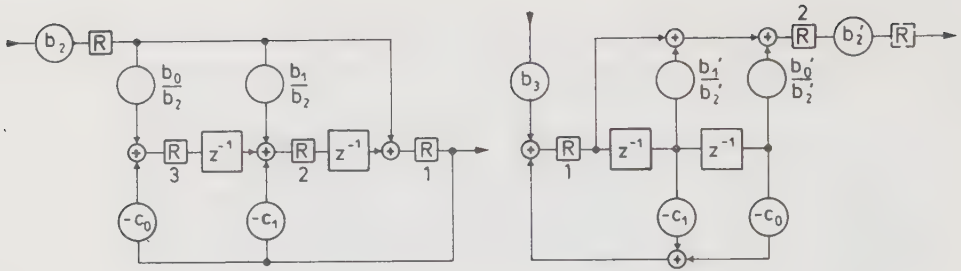
Tabelle 5.12: Ergebnisse verschiedener Optimierungsverfahren bei der grenzzyklenfreien Struktur (11)
Tiefpaß 8.Grades (Beispiel Nr.2)
• heuristische Lösung

Im vorliegenden Beispiel wurde mit der "dynamischen Programmierung" die bei fester Zuordnung auch unter Berücksichtigung aller Korrelationen bestmögliche Lösung erzielt. Bei gleicher notwendiger zusätzlicher Wortlänge $\Delta w = 6$ Bit liegt die Rauschleistung des "direkten Verfahrens" nur um 1.11 dB höher als die optimale Lösung. Entsprechende Ergebnisse wurden bei anderen Filtern erreicht.

Der Rechenzeitbedarf liegt bei Strukturen mit Abschneiden generell höher als bei ausschließlichem Runden. Ein Vergleich der verschiedenen Verfahren untereinander zeigt ebenfalls eine erhebliche Reduzierung der Rechenzeit beim "direkten Verfahren" mit fester Zuordnung.

5.3.4 Vergleich der Optimierungsergebnisse bei unterschiedlichen Realisierungen der Teilblöcke

Bei der Zusammenschaltung der Teilblöcke 2. Grades zu einer Kaskade ist die ohnehin schon große Anzahl der möglichen Realisierungen, wie sie in Abschnitt 5.2 behandelt wurden, noch durch die unterschiedlichen Realisierungen der Skalierung zu ergänzen. So ist es möglich, die Zählerverstärkungsfaktoren $b'_{2\lambda}$ in die übrigen Koeffizienten einzubeziehen (siehe Bild 5.53 und 5.54), oder diese durch getrennte Vorfaktoren bei den Strukturen 1. Art und durch Nachfaktoren bei den Strukturen 2. Art zu realisieren. Die Blockschaltbilder für die Grundstrukturen 1b-v und 2b-v werden im Bild 5.61 gezeigt. Darüberhinaus kann die Skalierung in Potenzen von zwei vorgenommen werden. Die Realisierung erfolgt dann durch einfache "shift"-Multiplikationen, die bei der Berechnung des Aufwands unberücksichtigt bleiben.



Struktur 1b-v

Struktur 2b-v

Bild 5.61: Grundstrukturen mit Vorfaktoren

Ein Vergleich aller möglichen Realisierungen ist auch hier im Rahmen der Abhandlung nicht möglich. Es werden daher nur die auf Grund der Voruntersuchungen in Abschnitt 5.2 für die gewählten Beispiele am günstigsten erscheinenden Grundstrukturen näher untersucht.

5.3.4.1 Verschiedene Realisierungen der 1. und 2. kanonischen Form bei Runden

Für Grundstrukturen der 1. und 2. kanonischen Form bei ausschließlichem Runden nach den Multiplikationen (1a, 2a) oder nach möglichen Additionen (1b, 2b) werden die Ergebnisse der "Dynamischen Programmierung" bei freier und fester (Abschnitt 5.3.2) Zuordnung der Pol- und Nullstellen in den Tabellen 5.13 und 5.14 zusammengestellt.

Die Gleichheit der Zählerfaktoren $b_2 = b_0$ wurde berücksichtigt. Bei der 2. kanonischen Form ergeben sich hieraus die beiden unterschiedlichen Lösungen. Werden nur 4 Multiplizierer realisiert, so ist nach Abschnitt 5.2.4.1 eine zusätzliche Übersteuerungsabfrage notwendig. Die Folge ist ein gegebenenfalls größeres Rauschen. Der Aufwand A_{SR} , der neben der Länge der Zustandsvariablen auch die Anzahl der Multiplikationen je Block berücksichtigt, wird jedoch geringer. Mit den Gln. (3.17), (3.24) und (4.102) erhält man für die Aufwandszahl A_{SR}

	Grundstruktur (Abbildung)	Zu- ord.	Anord- nung	R_1	$\angle w$	A_S	m_R	A_{SR}
Skalierung mit vollst. Multiplizierer	1 a (Bild 5.28)	frei	$\frac{34152}{34152}$	95,40	$\langle 3,07 \rangle = 4$	1,33	20	1,78
		fest	$\frac{34152}{34152}$	95,40	$\langle 3,07 \rangle = 4$	1,33	20	1,78
	1 b (Bild 5.26, 5.53) $b_2 = b_0$	frei	$\frac{34152}{34152}$	65,30	$\langle 2,80 \rangle = 3$	1,25	20	1,67
		fest	$\frac{34152}{34152}$	65,30	$\langle 2,80 \rangle = 3$	1,25	20	1,67
	1 a-v	frei	$\frac{52413}{42513}$	100,80	$\langle 3,11 \rangle = 4$	1,33	20	1,78
		fest	$\frac{34152}{34152}$	102,24	$\langle 3,12 \rangle = 4$	1,33	20	1,78
	1 b-v (Bild 5.61)	frei	$\frac{52413}{42513}$	78,39	$\langle 2,93 \rangle = 3$	1,25	20	1,67
		fest	$\frac{34152}{34152}$	80,47	$\langle 2,95 \rangle = 3$	1,25	20	1,67
Skalierung mit Potenzen von 2	1 a	frei	$\frac{34152}{35124}$	150,49	$\langle 3,40 \rangle = 4$	1,33	15	1,33
		fest	$\frac{35124}{35124}$	162,43	$\langle 3,45 \rangle = 4$	1,33	15	1,33
	1 b	frei	$\frac{34125}{35124}$	97,26	$\langle 3,09 \rangle = 4$	1,33	15	1,33
		fest	$\frac{35124}{35124}$	103,52	$\langle 3,13 \rangle = 4$	1,33	15	1,33
	1 a-v	frei	$\frac{23514}{32514}$	170,05	$\langle 3,50 \rangle = 4$	1,33	15	1,33
		fest	$\frac{32514}{32514}$	178,45	$\langle 3,53 \rangle = 4$	1,33	15	1,33
	1 b-v	frei	$\frac{52413}{42513}$	132,23	$\langle 3,31 \rangle = 4$	1,33	15	1,33
		fest	$\frac{32514}{32514}$	141,49	$\langle 3,36 \rangle = 4$	1,33	15	1,33

Tabelle 5.13: 1. kan. Form unterschiedliche Realisierung der Kaskade eines Tiefpasses 10. Grades
(Beispiel 4) $m_{Rmin} = 15$ $w_{min} = 12$
v = Realisierung mit Vorfaktor

	Grundstruktur (Abbildung)	Zu- ord.	Anord- nung	R_i	Δw	A_S	m_R	A_{SR}
Skalierung mit vollst. Multiplizierer	2 a	frei	$\frac{31524}{43152}$	141,22	$\langle 3,35 \rangle = 4$	1,33	21	1,87
	b_2, b_0 ge- meinsam	fest	$\frac{31524}{43152}$	141,22	$\langle 3,35 \rangle = 4$	1,33	21	1,87
	2 a	frei	$\frac{53124}{53412}$	106,67	$\langle 3,15 \rangle = 4$	1,33	26	2,31
	b_2, b_0 ge- trennt (Bild 5.28)	fest	$\frac{35124}{43512}$	108,50	$\langle 3,16 \rangle = 4$	1,33	26	2,31
	2 b	frei	$\frac{31524}{43152}$	36,9	$\langle 2,39 \rangle = 3$	1,25	21	1,75
	b_2, b_0 ge- meinsam (Bild 5.29)	fest	$\frac{31524}{43152}$	36,9	$\langle 2,39 \rangle = 3$	1,25	21	1,75
	2 b	frei	$\frac{35124}{43512}$	22,8	$\langle 2,04 \rangle = 3$	1,25	26	2,17
	b_2, b_0 ge- trennt (Bild 5.54)	fest	$\frac{35124}{43512}$	22,8	$\langle 2,04 \rangle = 3$	1,25	26	2,17
	2 a-v	frei	$\frac{54312}{43521}$	129,36	$\langle 3,29 \rangle = 4$	1,33	21	1,87
	2 b-v (Bild 5.62)	frei	$\frac{54213}{43521}$	49,90	$\langle 2,61 \rangle = 3$	1,25	21	1,75
Skalierung in Potenzen von 2	2 a	frei	$\frac{45123}{43512}$	151,54	$\langle 3,48 \rangle = 4$	1,33	15	1,33
	b_2, b_0 ge- trennt	fest	$\frac{25134}{42513}$	170,83	$\langle 3,49 \rangle = 4$	1,33	15	1,33
	2 b	frei	$\frac{12435}{34152}$	32,94	$\langle 2,31 \rangle = 3$	1,25	15	1,25
	b_2, b_0 ge- trennt	fest	$\frac{54123}{35412}$	36,35	$\langle 2,38 \rangle = 3$	1,25	15	1,25
	2 a-v	frei	$\frac{54132}{43512}$	199,90	$\langle 3,61 \rangle = 4$	1,33	15	1,33
	2 b-v	frei	$\frac{54213}{43521}$	81,81	$\langle 2,96 \rangle = 3$	1,25	15	1,25

Tabelle 5.14: 2. kan. Form, unterschiedliche Realisierung der Kaskade eines Tiefpasses 10. Grades
(Beispiel Nr.4) $m_{Rmin} = 15$ $w_{min} = 12$
 v = Realisierung mit Vorfaktor

$$A_{SR} = A_S \cdot A_R = \left(1 + \frac{\Delta w}{w_{\min}} \right) \cdot \frac{m_R}{m_{R\min}} \quad (5.101)$$

Vergleicht man nun die Aufwandszahlen für das angegebene Beispiel, so sieht man, daß sich trotz höheren Rauschens die Einsparung von Multiplikationen stets günstig auf den Aufwand auswirkt. Die in diesem Sinne günstigste Realisierung wird mit der 2. kanonischen Form, bei 2er-Potenz-Skalierung, getrennter "Shift"-Multiplikation der Faktoren b_2 und b_0 sowie einer einzigen Rundungsstelle je Block und einer zusätzlichen am Ausgang des Filters erreicht. Wird jedoch die Skalierung mit vollständigen Multiplizierern durchgeführt, so schneidet die 1. kanonische Form (1b) etwas besser ab.

Die Strukturen mit einer getrennten Vorfaktorrealisierung erbringen in dem vorliegenden Beispiel bei gleicher Aufwandszahl ein geringfügig größeres Rauschen. Eine allgemeine Aussage über die Realisierung mit Vorfaktoren darf jedoch hieraus nicht gemacht werden.

So hat z.B. bei der Optimierung der Bandsperre 22. Grades [54] (Anhang: Beispiel 14) die Struktur mit Vorfaktor einen günstigeren Wert erbracht. Die Ergebnisse der Struktur 1b und 1b-v werden in Tab. 5.15 mit der von Avenhaus in [6] angegebenen Anordnung verglichen.

Durch die in Abschnitt 5.3.3.2 beschriebene Optimierung konnte das Rauschverhalten weiter verbessert werden, wenngleich auf Grund des ganzzahligen Δw keine Verbesserung der Aufwandszahl möglich ist.

	Struktur	Anordnung	R_i	Δw
optimiert mit $S_{\max} = 100$	1 b	5 8 2 10 6 3 9 4 11 1 7	4196	<5,56> = 6
	1 b-v	5 8 2 10 6 3 9 4 11 1 7	2889	<5,28> = 6
Avenhaus [6]	1 b	6 9 3 5 1 8 11 4 7 10 2	4360	<5,57> = 6
	1 b-v	6 9 3 5 1 8 11 4 7 10 2	3025	<5,31> = 6

Tabelle 5.15: Bandsperre 22. Grades 1. kan. Form

Abschließend sei nochmals auf die geringen Unterschiede der mit fester Zuordnung optimierten Ergebnisse bezogen auf die optimalen Lösungen bei freier Zuordnung hingewiesen.

Einschränkend ist zu bemerken, daß bei den Strukturen 2a-v und 2b-v (mit Vorfaktoren) die "Dynamische Programmierung" mit fester Zuordnung der Hilfsblöcke nicht angewendet werden kann.

5.3.4.2 Vergleich weiterer Realisierungen

Neben der 1. und 2. kan. Form erscheinen auf Grund der Untersuchungen in Abschnitt 5.2.3 nach Bild 5.10 und 5.11 auch die Strukturen 4b (4. kan. nach Hwang [86]) und 6b (1C nach Jackson [71]) für ein Kaskadenfilter mit freiwählbaren Koeffizienten geeignet. Die Optimierungsergebnisse dieser Strukturen werden in Tab. 5.16 mit den entsprechenden der 1. und 2. kan. Form verglichen. Als Beispiel werden Tief- und Bandpässe 8. und 10. Grades verwendet. Bei allen vier Beispielen liegen die Aufwandszahlen der Strukturen 1b und 2b nahe beieinander. Die Rauschzahlen bei der Struktur 4b liegen in der Größenordnung der 1. kan. Form (ebenfalls 3 Rauschquellen je Block). Die Struktur 6b, nach Abschnitt 5.2.3.2 mit der 2. kan. Form gleichwertig, zeigt bei der Zusammenschaltung ein ungünstigeres Verhalten, da die Rundung am Ausgang des Blocks nicht in den nachfolgenden Block einbezogen werden kann. Da bei den Strukturen 4b und 6b für $b_2 = b_0$ kein Multiplizierer eingespart werden kann, liegen die Aufwandsmaße bei den angeführten Beispielen bei gleichem Δw höher.

Ein Vergleich der Rauschzahlen mit den in Bezug auf Grenzzyklen günstigen Realisierungen wird zusammen mit den Simulationsergebnissen im Abschnitt 6.3.2 durchgeführt.

Struktur		Anordnung	R_i	Δw	A_S	A_R^*	A_{SR}
TP 8. Grades Beispiel Nr. 2 $R_e=0,33$	1b	$\frac{2143}{3142}$	92,28	$\langle 3,06 \rangle = 4$	1,33	1,33	1,78
	2b	$\frac{2413}{3241}$	55,45	$\langle 2,69 \rangle = 3$	1,25	1,42	1,77
	4b	$\frac{4213}{3412}$	94,14	$\langle 3,07 \rangle = 4$	1,33	1,75	2,33
	6b	$\frac{4213}{3412}$	66,02	$\langle 2,82 \rangle = 3$	1,25	1,75	2,19
TP 10. Grades Beispiel Nr. 4 $R_e=0,34$	1b	$\frac{34152}{34152}$	65,30	$\langle 2,80 \rangle = 3$	1,25	1,33	1,67
	2b	$\frac{31524}{43152}$	36,98	$\langle 2,39 \rangle = 3$	1,25	1,40	1,75
	4b	$\frac{53214}{43512}$	62,59	$\langle 2,77 \rangle = 3$	1,25	1,73	2,16
	6b	$\frac{35214}{43512}$	44,72	$\langle 2,53 \rangle = 3$	1,25	1,73	2,16
BP 8. Grades Beispiel Nr. 12 $R_e=0,125$	1b	$\frac{2413}{2413}$	93,83	$\langle 3,19 \rangle = 4$	1,33	1,33	1,78
	2b	$\frac{4312}{2431}$	44,69	$\langle 2,65 \rangle = 3$	1,25	1,42	1,77
	4b	$\frac{3421}{2431}$	100,13	$\langle 3,24 \rangle = 4$	1,33	1,75	2,33
	6b	$\frac{3412}{2431}$	70,47	$\langle 2,98 \rangle = 3$	1,25	1,75	2,19
BP 10. Grades Beispiel Nr. 13 $R_e=0,138$	1b	$\frac{31542}{31542}$	88,0	$\langle 3,14 \rangle = 4$	1,33	1,33	1,78
	2b	$\frac{14253}{31425}$	34,11	$\langle 2,45 \rangle = 3$	1,25	1,40	1,75
	4b	$\frac{21453}{31254}$	90,42	$\langle 3,16 \rangle = 4$	1,33	1,73	2,30
	6b	$\frac{12453}{31254}$	65,21	$\langle 2,92 \rangle = 3$	1,25	1,73	2,16

Tabelle 5.16: Vergleich verschiedener Filter bei Realisierungen mit unterschiedlichen Grundstrukturen

* die Einbeziehung des Gesamtvorfaktors führt bei Filtern 8. und 10. Grades zu unterschiedlichen Werten A_R

5.3.5 Einfluß der Entwurfsparameter auf das Rauschverhalten einer Kaskade

Der Realisierung einer Übertragungsfunktion geht, wie in den Abschnitten 2 und 3 beschrieben, die Approximation des vorgegebenen Toleranzschemas voraus. Im allgemeinen wird bei unterschiedlichen Approximationen nicht nur die Koeffizientenempfindlichkeit beeinflusst, sondern auch das Rauschverhalten verändert. Beide Größen gehen in den schaltungstechnischen Realisierungsaufwand ein.

Im Abschnitt 3.4.2 wurden die Aufwandsmaße für eine Approximation des gleichen Toleranzschemas mit einem Cauer- und einem Tschebyscheff-Entwurf im Durchlaßbereich verglichen. Nach Tab. 3.3 erbrachte die Cauer-Lösung die günstigere Aufwandszahl. Durch das Einbeziehen des Aufwandsfaktors A_g zur Bewertung der Zustandsvariablen wird das Aufwandsverhältnis weiter zu Ungunsten des Tschebyscheff-Entwurfs verschlechtert.

Liegt die Art der Approximation für einen Frequenzgang fest, so bleibt der Einfluß der Filtergraderweiterung und die Abhängigkeit von dem Entwurfsparameter C nach Abschnitt 2.2.2 zu prüfen.

5.3.5.1 Einfluß der Filtergraderweiterung

Im Abschnitt 3.3.2 wurde der Koeffizientenaufwand in Abhängigkeit von der Filtergraderweiterung für verschiedene Toleranzschemata untersucht. Der Einfluß auf das Rauschverhalten wird am Beispiel des Tieffaßtoleranzschemas nach Bild 2.5a überprüft:

Für die in Tab. 3.1 zusammengestellten Filterentwürfe wird eine Optimierung der Anordnung durchgeführt. Dabei bleiben die Anordnungen der Filter 11. und 12. Grades

auf gegebenenfalls suboptimale Lösungen beschränkt. Für eine Realisierung mit Teilblöcken in der 1. kanonischen Form ohne Vorfaktoren mit einer minimalen Anzahl an Nichtlinearitäten (Bild 5.53) zeigt die Tab. 5.17 die Rauschzahlen.

n	R_i	Δw	A_S	A_R^*	A_N	A_C	A
8	92,28	$<3,06 = 4>$	1,33	1,33	1,0	2,29	4,07
9	66,63	$<2,82 = 3>$	1,25	1,36	1,125	1,29	2,47
10	65,30	$<2,80 = 3>$	1,25	1,33	1,25	1,0	2,08
11	63,61	$<2,78 = 3>$	1,25	1,35	1,377	1,0	2,32
12	71,45	$<2,87 = 3>$	1,25	1,33	1,50	1,0	2,50

Tabelle 5.17: Einfluß der Filtergraderweiterung
Cauer-Approx. des Toleranzschemas Bild 2.5a

* Mehraufwand bei ungeradem Filtergrad berücksichtigt

Die schaltungstechnische Bewertung wird entsprechend den Gln. (3.17), (3.22) - (3.24) und (4.102) durchgeführt. Zur Bestimmung von A_R wird von minimal 3 Multiplizierern beim Block 2. Grades und 2 beim 1. Grades ausgegangen.

Würde die Bewertung allein auf Grund der Rauschzahlen R_i durchgeführt, so würde die Lösung $n = 11$ den günstigsten Wert erbringen. Eine gleichzeitige Berücksichtigung der Filtergraderweiterung ergäbe mit $A_{SRN} = A_S A_R A_N = 1,91$ die günstigste Lösung bei $n = 9$. Berücksichtigt man jedoch auch den Koeffizientenaufwand, so ergibt sich die optimale Lösung mit $A = 2,08$ für $n = 10$.

Das durchgerechnete Beispiel macht deutlich, daß nur durch eine Gesamtbewertung aller Einflüsse eine brauchbare Aussage gemacht werden kann. Die im Prinzip notwendige Korrektur der

Werte A_C durch die Einbeziehung der Skalierungsfaktoren in die Zählerkoeffizienten bleibt hier unberücksichtigt. Bei einer Realisierung mit Grundblöcken in der 2. kanonischen Form erhält man ein entsprechendes Ergebnis.

5.3.5.2 Abhängigkeit von der Entwurfskonstanten C

Die Optimierung der Koeffizientenwortlänge und der Reihenfolge kann in den vorgestellten Programmen nicht gleichzeitig erfolgen. Besonders bei Grundstrukturen ohne getrennten Vorfaktor wird zur Festlegung der einzelnen Optimierungsphasen eine Untersuchung des Rauschverhaltens in Abhängigkeit von der Entwurfskonstanten C notwendig.

Als Beispiel werden die zur Untersuchung der Koeffizientenempfindlichkeit nach Bild 3.4 verwendeten Datensätze (Tiefpässe 10. Grades) in der Anordnung

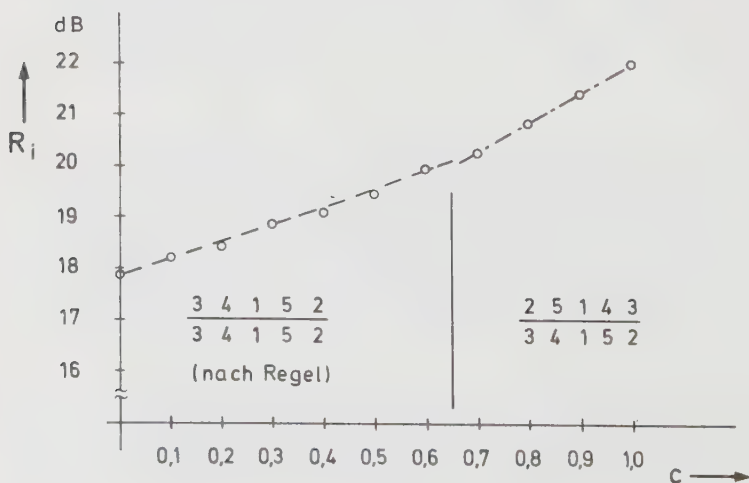


Bild 5.62: Rauschzahlen in Abhängigkeit von der Entwurfskonstanten c ; Tiefpässe 10. Grades, Struktur 1b.

der Teilfilter optimiert und die Rauschzahlen in Bild 5.62 in Abhängigkeit von der Entwurfskonstanten c aufgetragen.

Wie man sieht, steigen die Rauschzahlen mit größer werdender Konstanten c und somit stärkerer Ausnutzung des Toleranzschemas im Durchlaßbereich an. Die gefundene optimale Anordnung bleibt in einem großen, für die Koeffizientenoptimierung interessanten Bereich konstant.

Die in Abschnitt 5.3.2 beschriebene Zuordnungsregel für Pol- und Nullstellen wird im Bereich kleinerer Werte von c eingehalten.

Die Untersuchungen lassen daher den folgenden Arbeitsplan für eine schrittweise Optimierung bei Grundstrukturen ohne Vorfaktoren [101] sinnvoll erscheinen:

- In einer ersten Voroptimierung der Koeffizienten wird der Bereich der Konstanten c ungefähr festgelegt.
- Für diesen Vorentwurf wird die in Bezug auf das Rauschen optimale Anordnung mit einem der in Abschnitt 5.3.3 beschriebenen Verfahren aufgesucht.
- Unter Einbeziehung der Anordnung und Aussteuerung wird nach Abschnitt 3.2 die endgültige Optimierung der Koeffizientenwortlänge vorgenommen.

Nach der Auswahl einer günstigen Grundstruktur ist somit die Optimierung aller Kaskadenfilter in Bezug auf minimales Rauschen und kleine Koeffizientenwortlängen - d.h. minimalen Aufwand - mit den vorgestellten Programmen möglich.

6. Simulation digitaler Systeme

In den vorausgehenden Abschnitten wurden verschiedene Verfahren zur analytischen Untersuchung des Fehlerverhaltens digitaler Systeme mit endlicher Genauigkeit der Koeffizienten und Zustandsvariablen angegeben. Mit Hilfe von statistischen Modellen konnte die Rauschleistung am Ausgang eines Filters bestimmt und somit als Kriterium für eine Optimierung verwendet werden.

Zur Erprobung der in Bezug auf Übertragungsverhalten und Rausch-Nutzleistung optimierten Realisierungen können, soweit vorhanden, flexible einstellbare Filter (z.B. [39]) verwendet werden. Liegt jedoch für die zu testenden Strukturen keine entsprechende "Hard-ware"-Einrichtung vor, so wird eine Simulation des Systems vor dem endgültigen Aufbau notwendig.

Darüberhinaus hat sich die Simulation bei der Verbesserung bestehender und der Entwicklung neuer Fehlermodelle als ein nützliches Hilfsmittel erwiesen [7]. Die analytischen Rauschleistungsberechnungen der vorausgehenden Abschnitte werden durch die nachfolgenden Ergebnisse bestätigt.

Zur Erfüllung dieser Aufgaben benötigt man ein flexibles Programmsystem, das leicht zu handhaben ist und in Verbindung mit eingeführten - problemorientierten Programmiersprachen - verwendet werden kann.

6.1 Programm zur Simulation digitaler Systeme

Zur Simulation digitaler Systeme in Festkomma-Arithmetik auf Großrechnern wurden von Ullrich [102] und Hardner [103] Verfahren beschrieben. Die Nachbildung der verschiedenen Festkomma-Verknüpfungen erfolgt dabei vorwiegend durch Gleitkomma-Größen in einer höheren Programmiersprache. Hieraus

ergeben sich relativ lange Rechenzeiten.

Zur Durchführung der einleitend genannten Aufgaben wird jedoch ein flexibles Programmsystem benötigt, das der späteren Realisierung möglichst angepaßt ist und geringe Rechenzeiten für einen Simulationsschritt erfordert.

Für eine ökonomische Nachbildung der einzelnen "Hard-ware"-Schritte eignet sich besonders die maschinenorientierte Assembler-Sprache, da hierbei ein großer Teil der INTEGER-Maschinen-Befehle direkt für die Simulation verwendet werden kann. Prinzipiell kann die Simulationsaufgabe an jeder Rechenanlage gelöst werden. Die Verwendung eines Prozessrechners, wie er heute in vielen Labors installiert ist, bietet jedoch besondere Vorzüge, da der direkte Zugriff zur Rechenanlage und die Möglichkeit, periphere Geräte mit in die Simulation einzuplanen, der Aufgabenstellung sehr entgegen kommen. Das im weiteren verwendete Programmsystem SIMDIS ($\hat{=}$ Simulation digitaler Systeme) wurde speziell für die Prozessrechenanlage PDP-11/45 erstellt.

Für Großrechenanlagen geschriebene Simulationsprogramme verwenden in der Regel eine auf das Problem zugeschnittene Beschreibung des zu simulierenden Systems durch eine listenförmige, benutzerfreundliche Dateneingabe.

Nachteilig wirkt sich hierbei aus, daß alle möglichen Simulationsfunktionen gleichzeitig im Programm vorhanden sein müssen und so bei leistungsstarken Systemen ein sehr hoher Speicherbedarf erforderlich wird. In Verbindung mit der listenförmigen Eingabe kann eine Matrix-Beschreibung |46| des zu simulierenden Systems erfolgen. Da sich jedoch in der Regel sehr schwach besetzte Matrizen ergeben, muß zur Bestimmung des Matrix-Gleichungssystems ein im Vergleich zur eigentlichen Simulation hoher Rechenaufwand in Kauf genommen werden.

Das hier verwendete Programmsystem ist dagegen in einer Unterprogrammstruktur aufgebaut. Die Beschreibung bzw. Pro-

grammierung der zu simulierenden Systeme erfolgt durch Aneinanderreihen von vorprogrammierten ASSEMBLER-Routinen entsprechend dem Blockschaltbild des Systems. Dabei können die einzelnen Simulationsblöcke entweder innerhalb eines FORTRAN-Programms durch CALL-Unterprogramme aufgerufen werden, oder innerhalb einer ASSEMBLER-Routine ebenfalls als Unterprogramme aufgerufen oder zur Erhöhung der Arbeitsgeschwindigkeit als vorgefertigte MACRO-Programmteile eingebaut werden.

Für einfache Anwendungen liegen Haupt-Programme vor, die zusammen mit dem vom Benutzer zu schreibenden Unterprogramm zur Strukturbeschreibung, die Eingabe und Auswertung der Simulationsdaten übernehmen. Eine ausführliche Beschreibung des Programmsystems SIMDIS sowie Beispiele zur Programmierung werden in [51] gegeben. Hier soll die Anwendung von SIMDIS bei der Untersuchung des Rauschverhaltens von digitalen Filtern näher behandelt werden.

6.2 Strukturuntersuchungen mittels Simulation

Eine Untersuchung der unterschiedlichen, im Abschnitt 5 behandelten Realisierungen, ist mit Hilfe der Simulation möglich. Die Ergebnisse dienen einerseits zur Bestätigung der analytischen Modelle und liefern andererseits eine Möglichkeit, auch Strukturen und Arithmetiken zu vergleichen, für die noch keine ausreichenden Modelle vorliegen.

Das Fehlerverhalten eines digitalen Systems hängt im allgemeinen vom Eingangssignal ab. So führt das betragsmäßige Abschneiden (Abschnitt 4.2) zu einem mit dem Signal korrelierten Fehleranteil und setzt somit eine genaue Kenntnis des Eingangssignals und seiner statistischen Parameter voraus. Die Untersuchung mit verschiedenen Eingangssignalen ermöglicht Aussagen über die Anwendbarkeit der analytischen Fehlermodelle.

Für den Betrieb der digitalen Systeme stehen daher die folgenden Signalquellen zur Verfügung:

a) "Soft-ware" Rauschgenerator

gleichverteilte Pseudo-Zufallsfolge $\{u\}$, $-1 \leq u \leq 1$,
in Festkomma-Darstellung mit einer Quantisierung $Q=2^{-15}$

b) "Hard-ware"-Rauschgenerator |104|

Der über ein "Interface" an die Rechenanlage an-
geschlossene externe digitale Rauschgenerator¹⁾ er-
möglicht die Eingangsfolgen

- gleichverteiltes oder gaußisches Pseudo-
Zufallssignal
- gleichverteiltes oder gaußisches echtes
Rauschsignal .

Eine Steuerung der Varianz ist vom Rechner aus
möglich. Die Quantisierung des Rauschsignals be-
trägt ebenfalls $Q = 2^{-15}$. Der Rauschgenerator
selbst liefert ein "weißes" Spektrum. Eine spek-
trale Bewertung ist durch Zwischenschalten eines
externen digitalen Filters oder durch Simulation
des Filters im Rechner stets möglich.

c) "Hard-ware"-Sinus-Generator |104|

Neben dem stets möglichen Aufruf einer im Rechner
implementierten Sinus- oder Cosinusfunktion²⁾
kann ein "Hard-ware"-Sinusgenerator¹⁾ an die Simu-
lation angeschlossen werden. Die Steuerung des
Generators kann ebenfalls vom Rechner erfolgen
($\Omega \in [0, \pi]$). Der Vorteil liegt in einer kurzen
Aufrufzeit der Eingangsfolge. Eine Umrechnung in
die Festkomma-Darstellung entfällt.

d) Allgemeines Eingangssignal

Die vorgenannten Signalquellen stellen häufig
verwendete Testsignale für die Simulation dar

-
- 1) FORTRAN-Aufrufzeit für einen einzelnen Signalwert $t \geq 23.7 \mu\text{sec}$
2) FORTRAN-Aufrufzeit für einen einzelnen Signalwert $t \geq 342.3 \mu\text{sec}$
(PDP 11/45)

und eignen sich besonders zu einer generellen Untersuchung der Systeme.

Für spezielle Anwendungen können natürlich geeignete Eingangssignale errechnet oder von weiteren Datenträgern über ein "Interface" zur Simulation herangezogen werden.

6.2.1 Bestimmung der Aussteuerung

Die Überprüfung der verschiedenen Übersteuerungspunkte innerhalb eines digitalen Systems kann mit Hilfe der in dem Simulationsprogramm implementierten Überlaufanzeige [51] erfolgen. Für die experimentelle Bestimmung der Aussteuerungsgrenze des Eingangssignals lassen sich verschiedene Wege beschreiten.

Ausgehend von der maximal möglichen Aussteuerung des Eingangssignals ($\alpha = 1$) wird das Test-Filter solange betrieben, bis ein Überlauf, gegebenenfalls auch erlaubter Überlauf (Abschnitt 4.3.1) auftritt. Nach vorgegebener Reduzierung des Eingangssignals $\alpha < 1$ und Zurücksetzen der Zustandsspeicher wird der Test wiederholt. Zur Bestimmung der wirklichen Aussteuerungsgrenze, bzw. des notwendigen Abschwächungsfaktors α_n wird entsprechend Bild 6.1 das Maximum aller Beträge der Zustandsvariablen an den j Übersteuerungspunkten gebildet und α_n

$$\alpha_n = \frac{\alpha}{\max_j \{ |x_j(k)|, |y(k)| \}} \quad (6.1)$$

berechnet.

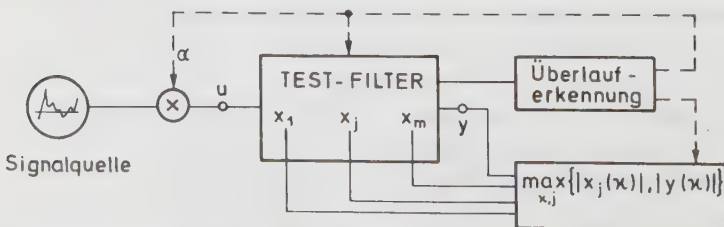


Bild 6.1: Bestimmung der Aussteuerung I

Bei Verwendung einer speziellen Fehlerüberwachung, die die erlaubten Bereichsüberschreitungen berücksichtigt, besteht die Möglichkeit einer direkten Angabe des maximalen Aussteuerungsfaktors. Die Anzahl der Überläufe $\ddot{u}$ in positiver (+) und in negativer (-) Richtung werden gespeichert. Ist am Übersteuerungspunkt die Gesamtzahl $\ddot{u} = 0$, so hat kein Überlauf stattgefunden.

Für $\ddot{u} \neq 0$ kann die Größe der Übersteuerung und somit der notwendige Abschwächungsfaktor aus dem fehlerhaften Signal $x_{\ddot{u}}$ berechnet werden.

Mit

$$x = x_{\ddot{u}} + 2 \ddot{u}$$

erhält man

$$\alpha_n = \frac{\alpha}{|x_{\ddot{u}} + 2 \ddot{u}|} \quad (6.2)$$

Der Aussteuerungstest wird auch hier solange fortgesetzt, bis keine Übersteuerung an den kritischen Punkten mehr auftritt.

6.2.2. Bestimmung der Gesamtrauschleistung

Die experimentelle Messung der Rauschleistung eines mit endlichen Zustandsspeichern realisierten digitalen Filters erfolgt entsprechend dem Blockschaltbild nach Bild 6.2. Um dabei jeglichen Einfluß der Koeffizientenquantisierung zu vermeiden, werden die Vergleichskoeffizienten des Norm-Filters aus den gerundeten Koeffizienten des Test-Filters bestimmt.

Die in Festkomma-Darstellung vorliegende Eingangsfolge $\{x\}$ wird sowohl von dem mit "SIMDIS"-programmierten Test-Filter, als auch von dem Norm-Filter in doppelt genauer Gleitkomma-Darstellung (PDP 11/45 54 Bit Mantisse) verarbeitet.

Die Auswertung der beiden Ausgangssignale erfolgt ebenfalls in doppeltgenauer Gleitkomma-Arithmetik. Man erhält

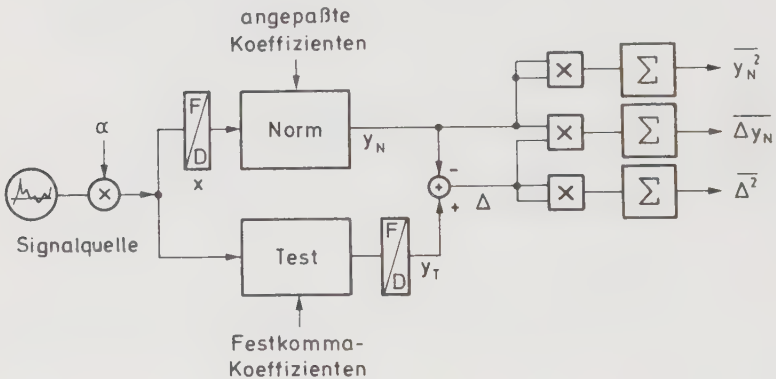


Bild 6.2: Rauschleistungsmessung

die Fehlerfolge $\{\Delta(k)\}$ am Ausgang des Vergleichers

$$\Delta(k) = y_T(k) - y_N(k) . \quad (6.3)$$

Die mittlere Fehlerleistung bestimmt sich zu

$$\overline{\Delta^2} = \overline{\Delta^2(k)} = \frac{1}{M} \sum_{k=1}^M \Delta^2(k) \quad (6.4)$$

Wird im Test-Filter eine Quantisierung auf Q_S durchgeführt, so kann man die Rauschzahl R_1 des Systems entsprechend Gl. (4.95) angeben.

$$R_1 = \frac{12}{Q_S} \overline{\Delta^2} . \quad (6.5)$$

6.2.3 Bestimmung der mit dem Signal linear unkorrelierten Rauschleistung

Aus den theoretischen Untersuchungen nach [7], [58] und den in Abschnitt 4.2 angegebenen Modellen zur Berechnung des Fehlers bei betragsmäßigem Abschneiden ist bekannt, daß das Fehlersignal $\Delta(k)$ teilweise mit dem Nutzsignal korreliert sein kann. Die Bestimmung des linear mit dem Signal korrelierten Fehleranteils, bzw. die Rauschzahl des linear unkorrelierten Fehlers R_{iO} kann ebenfalls mittels Simulation erfolgen.

Teilt man den Fehler in einen mit dem Signal linear unkorrelierten Anteil $\Delta_O(k)$ und einem mit K zum Signal proportionalen Anteil auf, so erhält man

$$\Delta(k) = \Delta_O(k) + K y_N(k) , \quad (6.6a)$$

und

$$\Delta_O(k) = \Delta(k) - K y_N(k) \quad (6.6b)$$

Der Proportionalitätsfaktor K wird nun nach [7] so bestimmt, daß der quadratische Mittelwert des Fehlers $\overline{\Delta_O^2(k)}$ ein Minimum wird, also

$$\frac{\partial}{\partial K} \overline{\Delta_O^2} \stackrel{!}{=} 0 \quad (6.7)$$

Durch Quadrieren erhält man

$$\overline{\Delta_O^2} = \overline{\Delta^2} - 2K \overline{\Delta y_N} + K^2 \overline{y_N^2} \quad (6.8)$$

und die Ableitung

$$\frac{\partial}{\partial K} \overline{\Delta_O^2} = -2 \overline{\Delta y_N} + 2K \overline{y_N^2} \stackrel{!}{=} 0$$

Hieraus ergibt sich der Proportionalitätsfaktor K zu

$$K = \frac{\overline{\Delta(k) \cdot y_N(k)}}{\overline{y_N^2(k)}} = \frac{\overline{\Delta y_N}}{\overline{y_N^2}} \quad (6.9)$$

und der der Leistung des linear unkorrelierten Fehlers entsprechende Mittelwert

$$\overline{\Delta_o^2(k)} = \overline{\Delta_o^2} = \overline{\Delta^2} - \frac{\overline{\Delta y_N^2}}{y_N^2} \quad (6.10)$$

Entsprechend ergibt sich die Rauschzahl R_{iO} des linear unkorrelierten Fehleranteils

$$R_{iO} = \frac{12}{Q_s^2} \overline{\Delta_o^2} \quad (6.11)$$

6.2.4 Notwendige Simulationslänge

In den vorausgehenden Abschnitten wurde ein Verfahren zur Bestimmung der mittleren Rauschleistung und somit der Rauschzahl eines digitalen Systems angegeben. Die Genauigkeit der Simulation hängt natürlich zwangsläufig von der Anzahl M der zur Bestimmung des Meßwertes durchgeführten Mittelungen ab. Einer beliebigen Erhöhung der Simulationsschritte steht jedoch der proportional steigende Rechenzeitbedarf entgegen.

Eine Aussage über die Güte des Meßwertes kann mit Hilfe des Vertrauensbereichs für den Schätzwert gemacht werden. Nach [105] erhält man für den 95%-Vertrauensbereich der Varianz σ^2 einer normalverteilten Grundgesamtheit mit dem Mittelwert $\mu = 0$

$$\frac{\overline{\Delta^2}}{1 + 1.96 \sqrt{\frac{2}{M-1}}} \lesssim \sigma^2 \lesssim \frac{\overline{\Delta^2}}{1 - 1.96 \sqrt{\frac{2}{M-1}}} \quad (6.12)$$

für $M > 150$.

Vergleicht man die Histogramme der Fehler $\Delta(k)$ für verschiedene Simulationslängen M , so sieht man, daß sich die Fehlerverteilung mit steigender Simulationslänge der Normalverteilung annähert und somit die Voraussetzung für Gl. (6.12) erfüllt.

Am Beispiel eines Blockes 2. Grades in der 1. kan. Form (siehe Bild 5.5) werden die Schätzwerte von R_i und deren 95%-Vertrauensbereiche gezeigt.

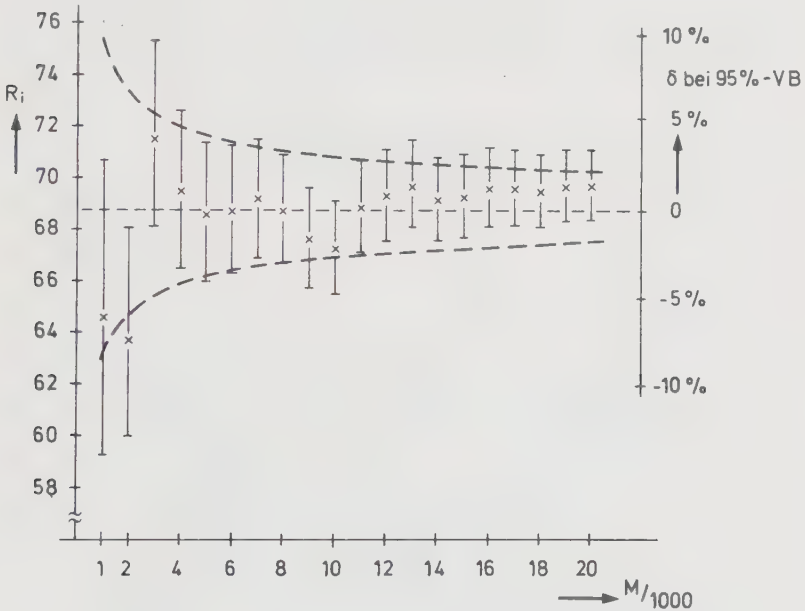


Bild 6.3: Messung der Rauschzahlen in Abhängigkeit von der Anzahl der Mittelungen M

Block 2. Grades $b_0=b_2=0.374$ $b_1= - 0.648$
 1.kan.Form $c_0 = 0.809$ $c_1= - 1.691$
 theoretischer Wert $R_i = 68,86$ (-----)

Gleichzeitig ist der theoretische Wert für R_i und der hierauf bezogene Vertrauensbereich δ in Abhängigkeit von der Anzahl der Simulationsschnitte eingezeichnet. Die Überprüfung der Histogramme zeigte, daß bei dem behandelten Beispiel für $M = 5000$ bereits mit guter Näherung eine Normal-Verteilung erreicht wird.

Diese Aussage läßt sich jedoch nicht verallgemeinern.

6.3 Ergebnisse verschiedener Strukturen

Das Rauschverhalten verschiedener Strukturen wurde im Abschnitt 5 anhand von theoretischen Modellen untersucht. Die so erhaltenen Ergebnisse wurden im Rahmen der Arbeit durch Simulation der Schaltungen nachgeprüft. Sie konnten mit sehr guter Übereinstimmung bestätigt werden.

Die Simulationsergebnisse einiger Beispiele werden im folgenden zum Vergleich zusammengestellt.

6.3.1 Grundstrukturen 2. Grades

Die unterschiedlichen Grundstrukturen von Blöcken 2. Grades wurden in Abschnitt 5.1 vorgestellt. Der Pol-Nullstellen-Test wird hier anhand der Simulation für die Strukturen 1 und 2 mit unterschiedlichen Arithmetiken sowie für die grenzzyklenfreien Strukturen 11 und 12 durchgeführt. Die entsprechenden Blockschaltbilder sind den Abschnitten 5.2.4 und 5.2.5 zu entnehmen. Zur Bestimmung der Rauschzahlen wurde über jeweils $M = 5000$ Fehlerwerte gemittelt. Als Filtereingangssignal wurde ein gleichverteiltes Rauschen verwendet. Die Ergebnisse der Simulation sind in den Bildern 6.4 und 6.5 zusammengestellt.

Neben dem generellen Rauschverhalten dieser Strukturen, das bereits im Abschnitt 5.2 diskutiert wurde, ist der Einbruch der Kurven für die Strukturen 1a, 1c, 2a, 2c bei $\psi_P = 90^\circ$ bemerkenswert. Hier wird der Koeffizient $c_1 = 0$. Somit entfällt in den Strukturen jeweils eine Rundungsquelle und die Rauschleistung am Ausgang vermindert sich entsprechend.

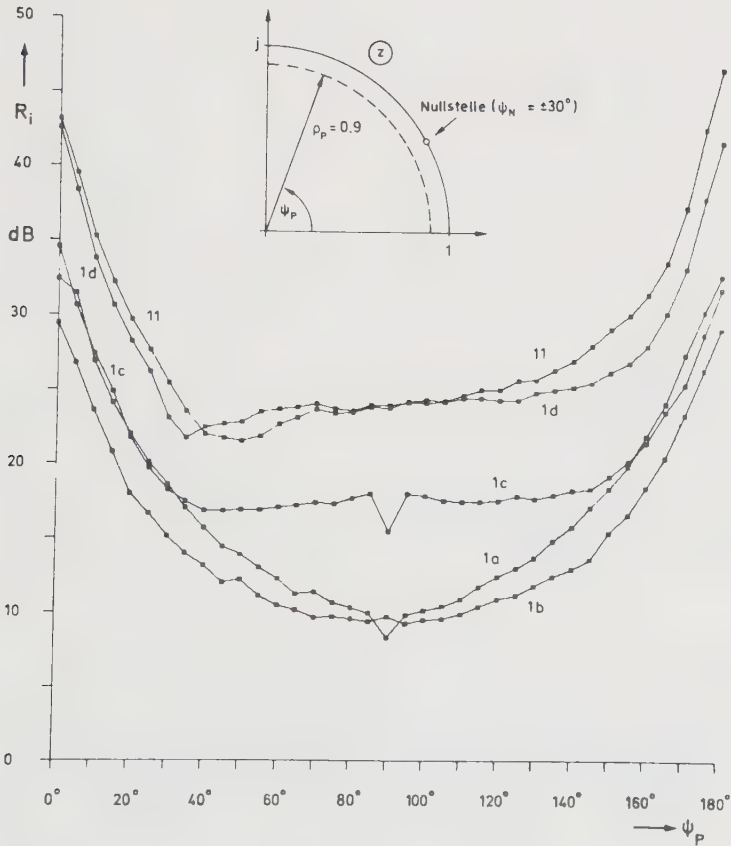


Bild 6.4: Pol-Nullstellen-Test für verschiedene Grundstrukturen der 1. Gruppe (Simulation)

Einen Vergleich der Gesamtrauschzahlen R_i und der mit dem Signal linear unkorrelierten Rauschzahlen R_{io} wird im Bild 6.6 für die Struktur 11 gezeigt. Wie man sieht, hängt der korrelierte Anteil stark von der Übertragungsfunktion

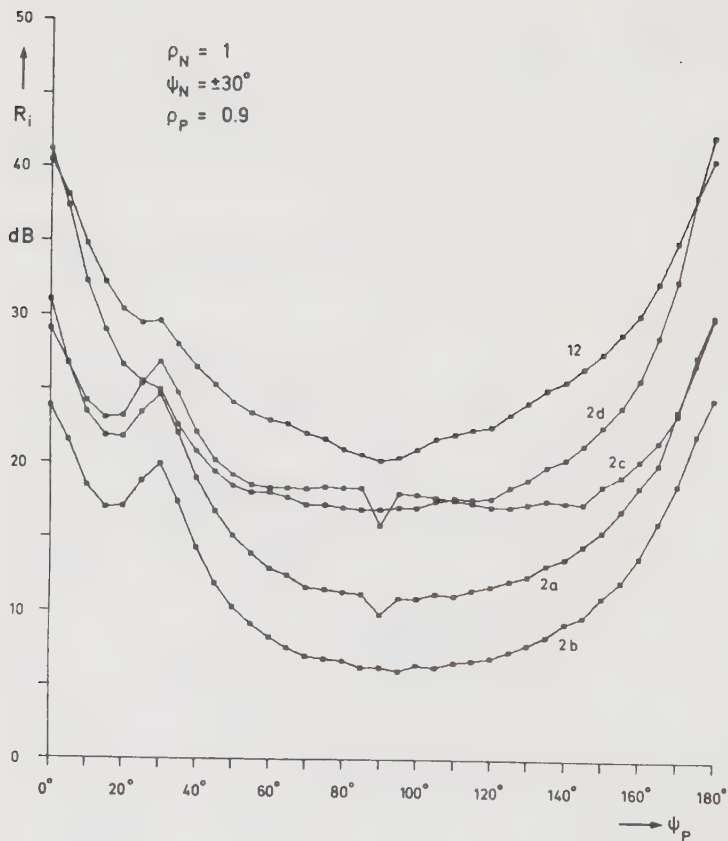


Bild 6.5: Pol-Nullstellen-Test für verschiedene Grundstrukturen der 2. Gruppe (Simulation)

ab und wird für Pole in der Nähe der Nullstelle verschwindend klein. Weitere Untersuchungen haben gezeigt, daß Strukturen mit ausschließlich Runden keinen korrelierten Fehleranteil aufweisen.

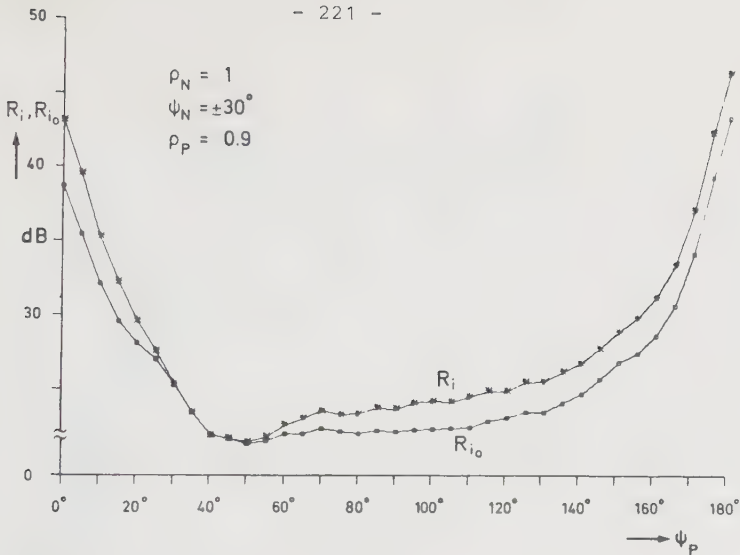


Bild 6.6: Rauschzahl des gesamten Fehlers und des linear unkorrelierten Anteils (Struktur 11)

Die gute Übereinstimmung der theoretischen Ergebnisse mit den simulierten wird in Bild 6.7 für die Strukturen 1b und 11 verdeutlicht. Die gemessenen Abweichungen sind kleiner 1 dB.

Der Einfluß der Polwinkel ψ_P bei konstantem Polradius ρ_P wurde in den vorausgehenden Darstellungen untersucht. Zur gleichzeitigen Betrachtung der Abhängigkeiten von Polwinkel- und -radius können auch hier die Linien gleicher Rauschzahlen aus einem Gitter von Meßpunkten über der komplexen Polebene bestimmt werden. Ergebnisse werden in Bild 6.8 für die Strukturen 1b und 11 gezeigt. Das schlechtere Rauschverhalten der grenzzyklenfreien Struktur wird über der ganzen Ebene anhand der eingetragenen Zahlenwerte sichtbar. Darüberhinaus erkennt man den rauschmindernden Einfluß der Nullstelle bei der Struktur 11.

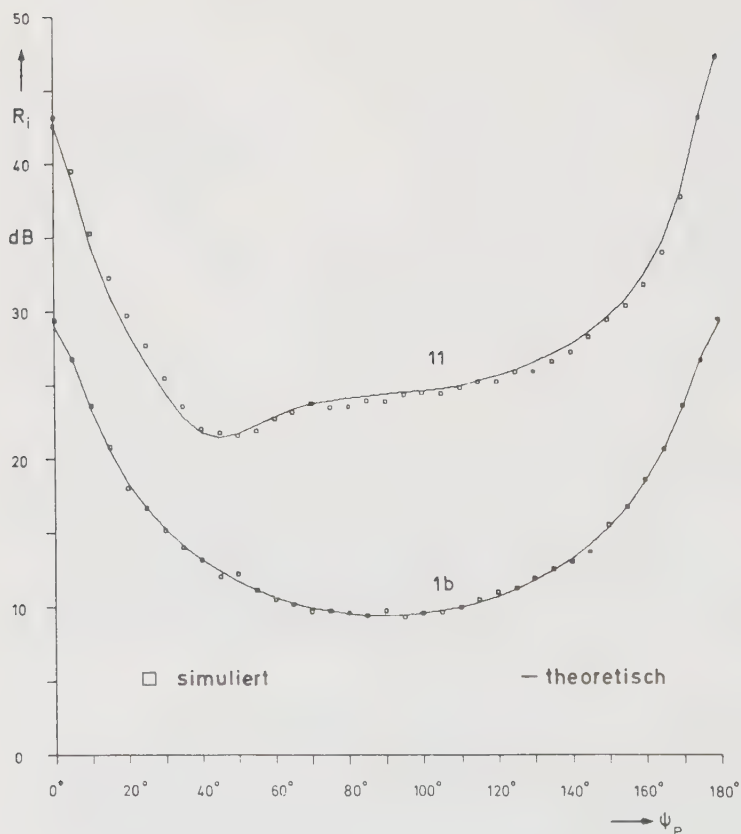


Bild 6.7: Vergleich der simulierten und der analytischen Ergebnisse

Das Rauschen der Strukturen mit betragsmäßigen Abschneidestellen hängt vom Eingangssignal ab. Das Fehlerverhalten der Struktur 11 wurde daher für Eingangssignale mit unterschiedlicher Verteilung untersucht. Die Meßergebnisse sind für einige Polwinkel ψ_p bei konstantem Polradius $\rho_p = 0.9$ und einer Nullstelle $\psi_N = \pm 30^\circ$ in der Tabelle 6.1 aufgelistet. Ein Vergleich der Rauschzahlen zeigt, daß bei den rauschartigen Eingangssignalen nur geringe Schwankungen auftreten. Eine prinzipielle Abweichung ist nicht zu erkennen.

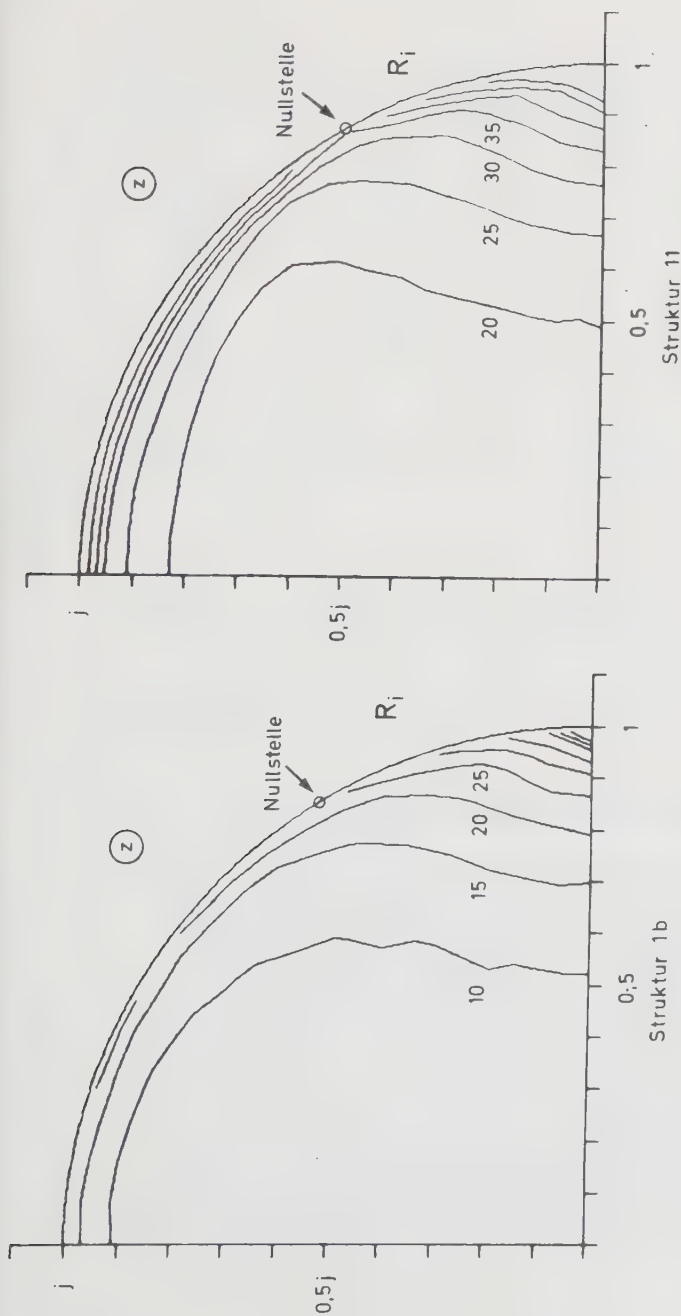


Bild 6.8: Linien konstanter Rauschzahlen R_i

ψ_p	gleichverteilt	normalverteilt		sinus $\Omega = \pi/3 \approx 60^\circ$
		$\sigma = 0,144$	$\sigma = 0,072$	
0°	43,10	42,65	42,22	31,37
30°	25,38	24,52	24,42	21,42
60°	22,70	22,90	22,69	<u>28,62</u>
90°	23,84	24,48	24,36	16,43
120°	25,12	25,33	25,44	15,69
150°	29,22	29,40	29,33	21,68
180°	46,86	46,36	46,22	29,42

Tabelle 6.1: Rauschzahlen in dB bei verschiedenen Eingangssignalen $\rho_p = 0,9$ $\psi_N = \pm 30^\circ$ $\rho_N = 1$

Anders verhält sich die Struktur bei sinusförmigem Eingangssignal. Wird die Schaltung mit einer Frequenz in der Nähe ihrer Eigenwerte erregt ($\psi_p = \pm 60^\circ$), so steigt die Rauschleistung durch die stärkere Bewertung des korrelierten Fehleranteils über die bei rauschartigem Eingangssignal an.

6.3.2 Das Rauschverhalten von Tiefpässen 8. und 10. Grades

Als Beispiele für die Untersuchung höhergradiger Filter in Kaskadenform werden die beiden Tiefpässe 8. Grades (Beispiel Nr.2) und 10. Grades (Beispiel Nr.4) verwendet. Aus der Vielzahl der möglichen Grundstrukturen werden die Struktur 1b und 2b, günstig in Bezug auf das Rauschen, 1c und 2d, günstig in Bezug auf das Grenzzyklen-Verhalten, sowie die grenzzyklenfreien Strukturen 11 und 13 ausgewählt. Für die Strukturen 1b und 2b wurde eine optimale Zuordnung, Reihenfolge und Skalierung mit dem in Abschnitt 5.3 beschriebenen Optimierungsprogramm errechnet. Die Strukturen mit betragsmäßigem Abschneiden und kontrolliertem Runden wurden für das Filter 8. Grades durch vollständigen

	Struktur	Optimierung mit analytischem Modell	Simulation							
			Anordnung Rauschzahl R_i	gesamtes Rauschen			linear unkorreliertes Rauschen			
				R_i	R_i/dB	$\Delta w/Bit$	R_{i0}	R_{i0}/dB	$\Delta w/Bit$	$\frac{R_i - R_{i0}}{R_i}$
Tiefpaß 8. Grades	1 b	$\frac{2143}{3142}$	92,28	95,44	19,80	3,08 = 4	95,44	19,80	3,08 = 4	0
	1 c	$\frac{1342}{2341}$	405,53	377,80	25,77	4,07 = 5	308,76	24,89	3,93 = 4	18,27%
	11	$\frac{1234}{2341}$	2713,85	2762,75	34,41	5,51 = 6	1684,31	32,26	5,15 = 6	39,04%
	2 b	$\frac{4123}{3412}$	32,69	32,90	15,17	2,31 = 3	32,90	15,17	2,31 = 4	0
	2 d	$\frac{3124}{3412}$	414,65	450,64	26,54	4,20 = 5	157,50	21,97	3,44 = 4	65,05
	13	$\frac{1324}{4231}$	1830,25	1707,01	32,32	5,16 = 6	1682,30	32,25	5,15 = 6	1,45%
Tiefpaß 10. Grades	1 b	$\frac{34152}{34152}$	65,30	67,92	18,32	2,83 = 3	67,92	18,32	2,83 = 3	0
	1 c	$\frac{42513}{42513}$	360,34	327,72	25,15	3,96 = 4	260,17	24,15	3,80 = 4	20,61%
	11	$\frac{24351}{24351}$	3001,53	3013,25	34,79	5,56 = 6	1657,71	32,20	5,13 = 6	44,99%
	2 b	$\frac{35124}{43512}$	22,77	23,02	13,62	2,05 = 3	23,02	13,62	2,05 = 3	0
	2 d	$\frac{34125}{53412}$	404,53	440,48	26,44	4,18 = 5	141,99	21,52	3,36 = 4	67,76%
	13	$\frac{23415}{52341}$	1359,91	1233,10	30,91	4,99 = 5	1174,62	30,68	4,89 = 5	4,73%

Tabelle 6.2: Ergebnisse eines Tiefpasses 8. und 10. Grades

dige Enumeration, die des Filters 10. Grades durch eine, auf eine feste Zuordnung eingeschränkte Enumeration optimiert. Die Ergebnisse der analytischen Untersuchung und der Simulation sind in Tab. 6.2 zusammengestellt. Die Ziffern in Spalte 2 geben die Anordnung der Teilblöcke an. Die Simulation wurde dabei mit gleichverteiltem Rauschen und $M = 30\,000$ Mittelungen durchgeführt. Ein Vergleich der so erhaltenen Werte mit den theoretischen zeigt auch hier eine gute Übereinstimmung. Neben den Rauschzahlen wurden auch die notwendigen zusätzlichen Wortlängen Δw für die verschiedenen Realisierungen bestimmt. Für beide Filter erweist sich wiederum die Struktur 2b als besonders günstig. Ein Vergleich des Gesamtrauschens mit dem linear unkorrelierten Rauschanteil zeigt, daß nur für Strukturen mit ausschließlich Runden der linear korrelierte Anteil verschwindet. Eine Reihe weiterer Strukturen mit teilweise anderen Anordnungen der Zähler- und Nennerblöcke wurden bereits in [106] untersucht.

Eine weitere Prüfung der Strukturen wurde mit sinusförmigen Eingangssignalen durchgeführt. In Bild 6.9 werden die Rauschzahlen des gesamten und des unkorrelierten Fehlers für die Strukturen 1b und 11 in Abhängigkeit von der Eingangsfrequenz aufgetragen. Ein deutliches Ansteigen des Rauschens in der Nähe der Polfrequenzen bei der Struktur mit betragsmäßigem Abschneiden (11) wird sichtbar. Die Struktur 1b hingegen weist ein nahezu unabhängiges Rauschverhalten auf. Durch diese Messungen konnte das theoretische Modell zur Berechnung des korrelierten Fehlers bei sinusförmigem Eingangssignal (Abschnitt 4.2.3.2) bestätigt werden. In Tabelle 6.3 sind theoretische und simulierte Werte gegenübergestellt.

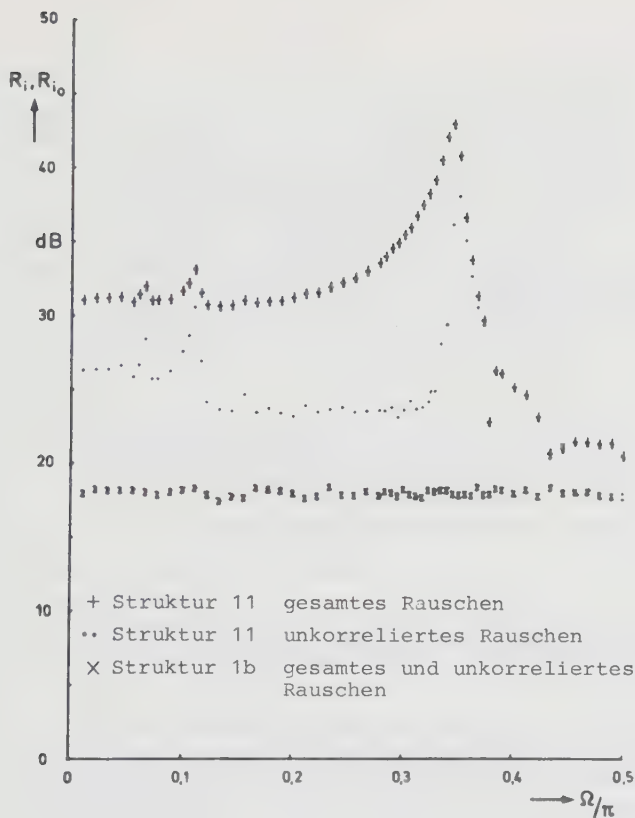


Bild 6.9: Rauschverhalten eines Tiefpasses 10. Grades bei sinusförmigem Eingangssignal

Eingangsfrequenz	0.1	0.15	0.2	0.25	0.3	0.35	0.4	Ω/π
simuliert	31.4	30.6	30.6	32.3	36.5	40.1	23.3	dB
theoretisch	31.83	31.19	31.37	32.63	35.63	40.78	26.02	dB

Tab. 6.3: Vergleich simulierter und theoretischer Werte bei sinusförmigem Eingangssignal.

7. Zusammenfassung

In der vorliegenden Arbeit werden die verschiedenen Einflüsse einer endlichen Koeffizienten- und Zustandsvariablen-Wortlänge bei digitalen Filtern mit unterschiedlichen Strukturen und Arithmetiken untersucht. Zum Vergleich verschiedener Filterentwürfe, die ein vorgegebenes Toleranzschema einhalten, werden schaltungstechnische Aufwandsmaße in vereinfachter Form definiert.

Speziell für die Kaskadenschaltung von Teilfiltern zweiten Grades in der ersten oder zweiten kanonischen Form wird gezeigt, daß durch eine Optimierung des Entwurfs eine Minimisierung des Koeffizientenaufwands möglich ist.

Die Untersuchung des Betriebsfehlerverhaltens verschiedener Grundstrukturen mit zum Teil unterschiedlichen Verkürzungsarithmetiken macht deutlich, daß für einen allgemeinen Block 2. Grades besonders Strukturen geeignet sind, die durch Zusammenfassung ungerundeter Teilprodukte eine Reduzierung der Rauschquellen ermöglichen. In Bezug auf die Rauschleistung erweist sich die ausschließliche Verwendung von "Runden" als vorteilhaft gegenüber anderen Arithmetiken.

Neben der Auswahl der Grundstrukturen ist die Optimierung der Anordnung der Zähler- und Nennerblöcke innerhalb einer Kaskade und die bei Festkomma-Arithmetik stets notwendige Skalierung der Teilfilter von entscheidendem Einfluß auf das Rausch-Signalleistungsverhältnis und somit auf den Realisierungsaufwand. Der Vergleich verschiedener Optimierungsverfahren zeigt, daß mit dem vorgestellten heuristischen Verfahren sehr gute Erfolge bei geringen Rechenzeiten erzielt werden können.

Ein abschließender Vergleich der optimalen Kaskadenschaltungen mit den am geeignetsten erscheinenden Grundblöcken bestätigt die Ergebnisse der Teilfilteruntersuchungen. Bei einer Betrachtung des Gesamtaufwands wird deutlich, daß nicht zwangsläufig der minimal mögliche Filtergrad die günstigste Lösung erbringt. Wohl aber erweisen sich im allgemeinen Strukturen mit einer minimalen Anzahl an vollständigen Multiplikationen, auch bei geringfügig größerem Rauschen, als besonders vorteilhaft. So erhält man für das angegebene Tiefpaß-Toleranzschema die günstigste Lösung mit einem Filter 10. Grades mit Grundblöcken in der 2. kanonischen Form und Zweier-Potenz-Skalierung.

Für ein freiprogrammierbares Kaskadenfilter erscheint neben der 2. kanonischen Form die Struktur 10 mit rückgekoppelten Speichern und inneren Skalierern wegen ihrer günstigen Rauscheigenschaften für Pole und Nullstellen in der Gegend von $z = \pm 1$ geeignet.

Die Erprobung der Filterentwürfe mit einem speziell für Festkomma-Arithmetik erstellten Simulationsprogramm bestätigt die Ergebnisse der analytischen Fehlermodelle.

LITERATURVERZEICHNIS

- |1| Schüßler, H.W.: Digitale Systeme zur Signalverarbeitung. Springer, Berlin 1973
- |2| Gold, B.; Rader, Ch.M.: Digital processing of signals. McGraw Hill, Inc., New York, 1969
- |3| Oppenheim, A.V.; Schafer, R.W.: Digital signal processing. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1975
- |4| Rabiner, L.R.; Gold, B.: Theory and application of digital signal processing. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1975
- |5| Jury, E.I.: Theory and application of the z-transform method. John Wiley & Sons, Inc., New York, 1964
- |6| Avenhaus, E.: Zum Entwurf digitaler Filter mit minimaler Speicherwortlänge für Koeffizienten und Zustandsgrößen. Ausgewählte Arbeiten über Nachrichtensysteme Nr. 13, hrsg. von W. Schüßler, Uni.Erlangen-Nürnberg, 1969
- |7| Eckhardt, B.: Untersuchungen des Multipliziererefehlers in digitalen Filtern. Ausgewählte Arbeiten über Nachrichtensysteme Nr. 21, hrsg. von W. Schüßler, Uni. Erlangen-Nürnberg, 1975
- |8| Deczky, A.G.: Synthesis of recursive digital filters using the minimum p-error criterion. IEEE Trans., Bd. AU-20, 1972, S. 257 - 263
- |9| Bandler, J.W.; Bardakjian, B.L.: Least pth optimization of recursive digital filters. IEEE Trans., Bd. AU-21, 1973, S. 460 - 470
- |10| Deczky, A.G.: Equiripple and minimax (chebyshev)-approximation for recursive digital filters. IEEE Trans., Bd. ASSP-22, 1974, S. 98 - 111
- |11| Rabiner, L.R.; Graham, N.Y.; Helms, H.D.: Linear programming design of IIR digital filters with arbitrary magnitude functions. IEEE Trans., Bd. ASSP-22, 1974, S. 117 - 123
- |12| Thajchayapong, P.; Rayner, P.J.W.: Design of recursive digital filters approximating to arbitrary prescribed magnitude response. Electr. Letters, Bd. 10, 1974, S. 159 - 161
- |13| Dudgeon, D.E.: Recursive filter design using differential correction. IEEE Trans., Bd. ASSP-22, 1974, S. 443- 448
- |14| Dubois, H.; Leich, H.: On the approximation problem of recursive digital filters with arbitrary attenuation curve in the pass-band and the stop-band. IEEE Trans., Bd. ASSP-23, 1975, S. 202 - 207

- [15] Burrus, C.S.; Parks, T.W.: Time domain design of recursive digital filters. IEEE Trans., Bd. AU-18, 1970, S. 137 - 141
- [16] Chao, K.-S.; Lu, K.-S.: On sequential refinement schemes for recursive digital filter design. IEEE Trans., Bd. CT-20, 1973, S. 396 - 401
- [17] Kikkawa, S.: Time-domain design of recursive filters for discrete inputs. Electronics and Communications in Japan, Bd. 56-A, 1973, S. 18 - 25
- [18] Steiglitz, K.: Designing short-word recursive digital filters. Proc. of the 9th Allerton Conference on Circuit and System Theory, 1971, S. 778 - 788
- [19] Suk, M.; Mitra, S.K.: Computer-aided design of digital filters with finite word lengths. IEEE Trans., Bd. AU-20, 1972, S. 356 - 363
- [20] Charalambous, Ch.; Best, M.J.: Optimization of recursive digital filters with finite word lengths. IEEE Trans., Bd. ASSP-22, 1974, S. 424 - 431
- [21] Bandler, J.W.; Bardakjian, B.L.; Chen, J.H.K.: Design of recursive digital filters with optimum word length coefficients. Computer Aided Design, Bd. 7, 1975, S. 151 - 156
- [22] Avenhaus, E.; Schüßler, H.W.: On the approximation problem in the design of digital filters with limited wordlength. AEÜ, Bd. 24, 1970, S. 571 - 572
- [23] Avenhaus, E.: On the design of digital filters with coefficients of limited wordlength. IEEE Trans., Bd. AU-20, 1972, S. 206 - 212
- [24] Boite, R.; Dubois, H.; Leich, H.: Optimization of digital filters in the discrete space of coefficients. Electr. Letters, Bd. 10, 1974, S. 179 - 180
- [25] Callahan, A.C.: A simplified design procedure for cascade form low-pass digital filters. IEEE Trans., Bd. ASSP-22, 1974, S. 381 - 383
- [26] Schüßler, H.W.: On the approximation problem in the design of digital filters. Proc. of the 5th Annual Princeton Conference on Information Science and Systems, 1971, S. 54 - 63
- [27] Sablatash, M.: The state of the art in approximation techniques for digital filter design. 1971 Symp. Digital Filtering, Imperial College of Science and Technology, London, Aug. 31 - Sept. 2, 1971
- [28] Cauer, W.: Theorie der linearen Wechselstromschaltungen. 2. Aufl., Akademie-Verlag, Berlin, 1954

- [29] Saal, R.; Ulbrich, E.: On the design of filters by synthesis. IRE Trans., Bd. CT-5, 1958, S. 284 - 327
- [30] Unbehauen, R.: Synthese elektrischer Netzwerke. München 1972
- [31] Rupprecht, W.: Netzwerksynthese, Entwurfstheorie linearer passiver und aktiver Zweipole und Vierpole. Berlin, 1972
- [32] Constantinides, A.G.: Spectral transformations for digital filters. Proc. IEE, Bd. 117, 1970, S. 1585 - 1590
- [33] Kuntz, W.; Schübler, W.; Winkelkemper, W.: Untersuchungen über Eigenschaften, Entwurf und Realisierung digitaler Filter. Ausgewählte Arbeiten über Nachrichtensysteme Nr. 10, hrsg. von W. Schübler, Uni. Erlangen-Nürnberg, 1969
- [34] Christian, E.; Eisenmann, E.: Filter design tables and graphs. John Wiley & Sons, Inc., New York, 1966
- [35] Winkelkemper, W.: Unsymmetrical bandpass and bandstop digital filters. Electron. Letters, Bd. 5, 1969, S. 585-586
- [36] Gadenz, R.N.; Temes, G.C.: Computational algorithm for the design of elliptic filters. Electron. Letters, Bd. 8, 1972, S. 323 - 324
- [37] Dehner, G.: On the design of digital Cauer filters with coefficients of limited wordlength. AEÜ, Bd. 29, 1975, S. 165 - 168
- [38] Eckhardt, B.; Winkelkemper, W.: Entwurf und Aufbau eines flexiblen rekursiven digitalen Filters. Tagungsband NTG-Tagung Signalverarbeitung 1973, hrsg. von H.W. Schübler, Uni. Erlangen-Nürnberg, 1973, S. 104 - 111
- [39] Eckhardt, B.; Winkelkemper, W.: Ein Beitrag zum Schaltungs-entwurf digitaler Filter für den Tonfrequenzbereich. NTZ, Bd. 27, No. 1, 1974, S. 2 - 10
- [40] Laxroix, A.: Zum Approximationsproblem bei digitalen Filtern mit Koeffizienten in Gleitkomma-Darstellung. DFG-Kolloquium, Digitale Systeme zur Signalverarbeitung, Uni. Erlangen-Nürnberg, 7. und 8. März 1974
- [41] Agarwal, R.C.; Burrus, C.S.: New structures for recursive digital filters having very low coefficient quantization and round-off noise. IEEE Trans., Bd. CAS-22, 1975, S. 921 - 927
- [42] Knowles, J.B.; Edwards, R.: Effect of finite-word-length computer in a sampled-data feedback system. Proc. IEE, Bd. 112, 1965, S. 1197 - 1207
- [43] Knowles, J.B.; Olcayto, E.M.: Coefficient accuracy and digital filter response. IEEE Trans., Bd. CT-15, 1968, S. 31 - 41

- [44] Rader, C.M.; Gold, B.: Effects of parameter quantization on the poles of digital filters. Proc. of the IEEE, Bd. 55, 1967, S. 688 - 689
- [45] Fettweis, A.: On the connection between multiplier word-length limitation and roundoff noise in digital filters. IEEE Trans., Bd. CT-19, 1972, S. 486 - 491
- [46] Crochiere, R.E.: Digital network theory and its application to the analysis and design of digital filters. Ph. D. dissertation, Dep. Elec. Eng., M.I.T., Cambridge, Mass., 1974
- [47] Crochiere, R.E.; Oppenheim, A.V.: Analysis of linear digital networks. Proc. of the IEEE, Bd. 63, 1975, S. 581 - 595
- [48] Belter, S.E.; Bass, S.C.: Computer-aided analysis and design of digital filters with arbitrary topology. IEEE Trans., Bd. CAS-22, 1975, S. 810 - 819
- [49] Hermeling, R.; Lückner, R.: Analyse und Empfindlichkeitsberechnung zeitdiskreter Systeme. Mitteilungen aus dem Fachbereich Nachrichtentechnik, Fachgebiet Übertragungstechnik der TH Darmstadt.
- [50] Fisz, M.: Wahrscheinlichkeitsrechnung und mathematische Statistik. VEB Deutscher Verlag der Wissenschaften, Berlin, 1966
- [51] Dehner, G.: Programme zum Entwurf und zur Simulation digitaler Filter. Ausgewählte Arbeiten der Nachrichtentechnik Nr. 24, hrsg. von H.W. Schüsler, Uni. Erlangen-Nürnberg, in Vorbereitung.
- [52] Schüsler, H.W.: Ein Gütemaß für die Realisierung digitaler Filter. DFG-Kolloquium, Bochum 1975
- [53] Heute, U.: Über Realisierungsprobleme bei nicht-rekursiven Digitalfiltern. Ausgewählte Arbeiten der Nachrichtentechnik Nr. 20, hrsg. von H.W. Schüsler, Uni. Erlangen-Nürnberg, 1975
- [54] Golden, R.M.; Kaiser, J.F.: Design of wideband sampled-data filters. Bell Syst. Tech. J., Bd. 43, 1964, S. 1533 - 1546
- [55] Knowles, J.B.; Edwards, R.: Autocorrelation function of the quantisation-error process for sinusoid. Electr. Letters, Bd. 4, 1968, S. 180 - 182
- [56] Hirano, K.; Ito, Y.; Ishikawa, K.; Mitra, S.K.: Statistical analysis of quantization errors in digital filters with finite wordlengths. Zur Veröffentlichung eingereicht.
- [57] Liu, B.; van Valkenburg, M.E.: On roundoff errors for fixed-point digital filters using sign-magnitude truncation. IEEE Trans., Bd. CT-19, 1972, S. 536 - 537

- [58] Claasen, T.A.C.M.; Mecklenbräuker, W.F.G.; Peek, J.B.H.: Quantization noise analysis for fixed-point digital filters using magnitude truncation for quantization. IEEE Trans., Bd. CAS-22, 1975, S. 887 - 895
- [59] Weinstein, C.; Oppenheim, A.V.: A comparison of roundoff noise in floating-point and fixed-point digital filter realization. Proc. of the IEEE, Bd. 57, S. 1181 - 1183
- [60] Lacroix, A.: Entwurf und Realisierung eines Sprachsynthesators durch ein digitales Filter. Dissertation am Fachbereich Nachrichtentechnik der Techn. Hochschule Darmstadt, 1974
- [61] Claasen, T.A.C.M.; Mecklenbräuker, W.F.G.; Peek, J.B.H.: Second-order digital filter with only one magnitude truncation quantizer and having practically no limit cycles. Electr. Lett., Bd. 9, 1973, S. 531 - 532
- [62] Meerkötter, K.; Wegener, W.: A new second-order digital filter without parasitic oscillations. AEÜ, Bd. 29, 1975, S. 312 - 314
- [63] Butterweck, H.J.: Suppression of parasitic oscillations in second-order digital filters by means of a controlled rounding arithmetic. AEÜ, Bd. 29, 1975, S. 371 - 374
- [64] Verkroost, G.; Butterweck, H.J.: Suppression of parasitic oscillations in wave digital filters and related structures by means of controlled rounding. AEÜ, Bd. 30, 1976, S. 181 - 186
- [65] Büttner, M.: Some experimental results concerning random-like noise and limit-cycles. NTZ, Bd. 28, 1975, S. 402 - 406
- [66] Büttner, M.: A novel approach to eliminate limit cycles in digital filters with a minimum increase in the quantization noise. ISCAS-Tagung, Apr. 1976, München, S. 291 - 294
- [67] Claasen, T.A.C.M.: Quantization-noise analysis of digital filters with controlled quantization. Electron. Lett., Bd. 12, 1976, S. 46 - 47
- [68] Fettweis, A.: Pseudopassivity, sensitivity and stability of wave digital filters. IEEE Trans., Bd. CT-19, 1972, S. 668 - 673
- [69] Eckhardt, B.: On the roundoff error of a multiplier. AEÜ, Bd. 29, 1975, S. 162 - 164
- [70] Bronstein, I.H.; Semendjajew, K.A.: Taschenbuch der Mathematik, Verlag Deutsch, Zürich-Frankfurt, 1966
- [71] Jackson, L.B.: An analysis of roundoff noise in digital filters. Sc.D.Thesis, Stevens Inst. of Technology, Hoboken, New Jersey, 1968

- [72] Jackson, L.B.: On the interaction of roundoff noise and dynamic range in digital filters. Bell Syst. Techn. J., Bd. 49, 1970, S. 159 - 184
- [73] Oppenheim, A.V.; Weinstein, C.J.: Effects of finite register length in digital filtering and the fast fourier transform. Proc. of the IEEE, Bd. 60, 1972, S. 957 - 976
- [74] Liu, B.: Effect of finite word length on the accuracy of digital filters - a review. Trans. IEEE, Bd. CT-18, 1971, S. 670 - 677
- [75] Eckhardt, B.; Winkelkemper, W.: Implementation of a second order digital filter section with stable overflow behaviour. NTZ, Bd. 26, 1973, S. 282 - 284
- [76] Avenhaus, E.: A proposal to find suitable canonical structures for the implementation of digital filters with small coefficient wordlength. NTZ, Bd. 25, 1972, S. 377 - 382
- [77] Parker, S.R.; Hess, S.: Canonic realizations of second-order digital filters due to finite precision arithmetic. IEEE Trans., Bd. CT-19, 1972, S. 410 - 413
- [78] Mitra, S.K.; Sherwood, R.: Canonic realizations of digital filters using the continued fraction expansion. IEEE Trans., Bd. AU-20, 1972, S. 185 - 194
- [79] Mitra, S.K.; Sherwood, R.: Digital ladder networks. IEEE Trans., Bd. AU-21, 1973, S. 30 - 36
- [80] Mitra, S.K.; Douglas, C.H.; Sherwood, R.J.: New methods of digital ladder realizations. IEEE Trans., Bd. AU-21, 1973, S. 485 - 491
- [81] Gray, A.H.; Markel, J.D.: Digital lattice and ladder filter synthesis. IEEE Trans., Bd. AU-21, 1973, S. 491 - 500
- [82] Gray, A.H.; Markel, J.D.: A normalized digital filter structure. IEEE Trans., Bd. ASSP-23, 1975, S. 268 - 277
- [83] Markel, J.D.; Gray, A.H.: Fixed-point implementation algorithms for a class of orthogonal polynomial filter structures. IEEE Trans., Bd. ASSP-23, 1975, S. 486 - 494
- [84] Szcupak, J.; Mitra, S.K.: Digital filter realization using successive multiplier-extraction approach. IEEE Trans., Bd. ASSP-23, 1975, S. 235 - 239
- [85] Hwang, S.Y.: Realization of canonical digital networks. IEEE Trans., Bd. ASSP-22, 1974, S. 27 - 39
- [86] Hwang, S.Y.: On element-values and realizability conditions of canonical digital networks. IEEE Trans., Bd. ASSP-22, 1974, S. 152 - 153
- [87] Bonzanigo, F.; Kinzl, A.: The sensitivity of the regular canonical digital filter structures. Florence Conference on Digital Signal Processing, Florence, Italy, Sept. 1975

- | 88 | Mitra, S.K.; Hirano, K.: Digital all-pass networks. IEEE Trans., Bd. CAS-21, 1974, S. 688 - 700
- | 89 | Gorille, I.: Digitale Simulation und Optimierung von Nachrichtensystemen. Ausgewählte Arbeiten der Nachrichtentechnik Nr. 14, hrsg. von H.W. Schüßler, Uni. Erlangen-Nürnberg, 1971
- | 90 | Jackson, L.B.: Roundoff-noise analysis for fixed-point digital filters realized in cascade or parallel form. IEEE Trans., Bd. AU-18, 1970, S. 107 - 122
- | 91 | Avenhaus, E.: Zur Realisierung digitaler Filter mit günstigem Nutz-Störsignalverhältnis, NTZ; Bd. 23, 1970, S. 217 - 219
- | 92 | Lee, W.S.: Optimization of digital filters for low roundoff noise. IEEE Trans., Bd. CAS-21, 1974, S. 424 - 431
- | 93 | Hwang, S.Y.: On optimization of cascade fixed-point digital filters. IEEE Trans., Bd. CS-21, 1974, S. 163 - 166
- | 94 | Lueder, E.; Hug, H.; Wolf, W.: Minimizing the round-off noise in digital filters by dynamic programming. Frequenz, Bd. 29, 1975, S. 211 - 214
- | 95 | Liu, B.; Peled, A.: Heuristic optimization of the cascade realization of fixed-point digital filters. IEEE Trans., Bd. ASSP-23, 1975, S. 464 - 473
- | 96 | Jackson, L.B.: Roundoff noise bounds derived from coefficient sensitivities for digital filters. Zur Veröffentlichung eingereicht
- | 97 | Dehner, G.: A contribution to the optimization of roundoff-noise in recursive digital filters. AEÜ, Bd. 29, 1975, S. 505 - 510
- | 98 | Chan, D.S.K.; Rabiner, L.B.: An algorithm for minimizing roundoff-noise in cascade realizations of finite impulse response digital filters. Bell Syst. Techn. J., Bd. 52, 1973, S. 347 - 385
- | 99 | Müller-Merbach, H.: Optimale Reihenfolgen. Springer, Berlin, 1970
- | 100 | Hoffmann, U.; Hofmann, H.: Einführung in die Optimierung, Verlag Chemie GmbH, Weinheim
- | 101 | Schüßler, H.W.: Persönliche Mitteilungen, 1974

- | 102 | Ullrich, U.: Simulation von Digitalfiltern mit Hilfe von Großrechnern. AEÜ, Bd. 28, 1974, S. 81 - 86
- | 103 | Harder, W.: Zeitanalyse mit digitaler Simulation. Nachrichtentechnik - Elektrotechnik, Bd. 24, 1974, S. 219 - 221
- | 104 | Büttner, M.; Eckhardt, B.; Oetken, G.: Meßtechnik in der seriellen digitalen Signalverarbeitung. Ausgewählte Arbeiten aus der Nachrichtentechnik Nr. 22, hrsg. von H.W. Schüßler, Uni.Erlangen-Nürnberg, 1976
- | 105 | Sachs, L.: Statistische Methoden, Springer Berlin 1970
- | 106 | Dehner, G.: On the noise behaviour of digital filter block of second order. AEÜ zur Veröffentlichung eingereicht

Anhang: Datensätze der Entwurfsbeispiele

1. Tiefpässe

Toleranzschema: $\Omega_D = \pi/3$ $\Omega_S = 1.18$
 $\delta_D = 0.02$ $\delta_S = 0.001$

1.1 Approximation nach CAUER $n_{\min} = 7.99$

1.1.1 minimaler Filtergrad, gleichmäßige Ausnutzung
bei ungerundeten Koeffizienten

1	n = 8	$\Delta n/n_{\min} \quad 0.0001, \quad C = 14.25, \quad c = 0.5$			
K = 0.00767				$w_c = \infty$	
v	b_{2v}	b_{1v}	b_{ov}	c_{1v}	c_{ov}
1	1.0	- 0.746022	1.0	- 0.960598	0.956259
2	1.0	- 0.585286	1.0	- 0.992710	0.837902
3	1.0	- 0.033073	1.0	- 1.093023	0.636708
4	1.0	1.466143	1.0	- 1.211321	0.418276
notwendige Wortlänge $w = 19$ Bit, $q_{b1} = q_{c1} = 17, \quad q_{co} = 18$					

1.1.2 minimaler Filtergrad, Koeffizienten optimiert

2	n = 8	$\Delta n/n_{\min} \quad 0.0001, \quad C = 14.23, \quad c = 0.265$			
K = 0.00767				$w_c = \infty$	
v	b_{2v}	b_{1v}	b_{ov}	c_{1v}	c_{ov}
1	1.0	- 0.746022	1.0	- 0.960560	0.956242
2	1.0	- 0.585286	1.0	- 0.992624	0.837841
3	1.0	- 0.033073	1.0	- 1.092835	0.636579
4	1.0	1.466143	1.0	- 1.211015	0.418085
K = 0.00767 $q_{b1} = q_{c1} = 14, \quad q_{co} = 15$				$w_c = 16$	
v	b_{2v}	b_{1v}	b_{ov}	c_{1v}	c_{ov}
1	1.0	- 0.746032	1.0	- 0.960571	0.956237
2	1.0	- 0.585266	1.0	- 0.992614	0.837829
3	1.0	- 0.033081	1.0	- 1.092834	0.636566
4	1.0	1.466125	1.0	- 1.210998	0.418090

1.1.3 erweiterter Filtergrad, Koeffizienten optimiert

3	n = 9	$\Delta n/n_{\min} = 0.125, \quad C = 8.48, \quad c = 0.082$			
	K = 0.00733				$w_c = \infty$
v	b_{2v}	b_{1v}	b_{0v}	c_{1v}	c_{0v}
1	1.0	- 0.749475	1.0	- 0.933959	0.951937
2	1.0	- 0.628782	1.0	- 0.928679	0.830434
3	1.0	- 0.255395	1.0	- 0.951027	0.634741
4	1.0	0.764367	1.0	- 0.988710	0.385192
5	1.0	1.0		- 0.504727	
	K = 0.00733		$q_{b1} = q_{c1} = q_{co} = 8$		$w_c = 9$
v	b_{2v}	b_{1v}	b_{0v}	c_{1v}	c_{0v}
1	1.0	- 0.750000	1.0	-0.933537	0.953125
2	1.0	- 0.628906	1.0	-0.929687	0.832031
3	1.0	- 0.253906	1.0	-0.949218	0.632812
4	1.0	+ 0.765625	1.0	-0.988281	0.386718
5	1.0	1.0		-0.503906	

1.1.4 erweiterter Filtergrad, Koeffizienten optimiert

4	n = 10	$\Delta n/n_{\min} = 0.25, \quad C = 5.09, \quad c = 0.086$			
	K = 0.00695				$w_c = \infty$
v	b_{2v}	b_{1v}	b_{0v}	c_{1v}	c_{0v}
1	1.0	- 0.751924	1.0	- 0.913424	0.952110
2	1.0	- 0.657665	1.0	- 0.889780	0.836577
3	1.0	- 0.388098	1.0	- 0.873145	0.659143
4	1.0	0.299157	1.0	- 0.857794	0.415948
5	1.0	1.636252	1.0	- 0.846440	0.210392
	K = 0.00689		$q_{b1} = 5, \quad q_{c1} = q_{co} = 6$		$w_c = 7$
v	b_{2v}	b_{1v}	b_{0v}	c_{1v}	c_{0v}
1	1.0	- 0.750000	1.0	- 0.906250	0.953125
2	1.0	- 0.656250	1.0	- 0.890625	0.843750
3	1.0	0.375000	1.0	- 0.875000	0.656250
4	1.0	0.312500	1.0	- 0.859375	0.421875
5	1.0	1.625000	1.0	- 0.843750	0.203125

1.1.5 erweiterter Filtergrad, Koeffizienten mit
Skalierung - optimiert

5	n = 10	$\Delta n/n_{\min} = 0.25, C = 4.56, c = 0.041$			
K = 1.0		$R_i = 63.11 \quad \Delta w > 2.79$			$w_c = \infty$
v	b_{2v}	b_{1v}	b_{0v}	c_{1v}	c_{0v}
1	0.369980	- 0.143188	0.369980	- 0.862173	0.654593
2	0.266816	- 0.080189	0.266816	- 0.839644	0.408360
3	0.570368	- 0.428496	0.570368	- 0.909783	0.951421
4	0.187074	0.306201	0.187074	- 0.822127	0.200269
5	0.707626	- 0.464819	0.707626	- 0.883718	0.834316
K = 1.0		q = 7			$w_c = 8$
v	b_{2v}	b_{1v}	b_{0v}	c_{1v}	c_{0v}
1	0.367187	- 0.140625	0.367187	- 0.859375	0.656250
2	0.265625	- 0.078125	0.265625	- 0.835937	0.406250
3	0.570312	- 0.429687	0.570312	- 0.906250	0.953125
4	0.187500	0.304687	0.187500	- 0.820312	0.203125
5	0.710937	- 0.468750	0.710937	- 0.882812	0.835937

1.1.6 erweiterter Filtergrad, Koeffizienten optimiert

6	n = 11	$\Delta n/n_{\min} = 0.377, C = 3.40, c = 0.114$			
K = 0.006197					$w_c = \infty$
v	b_{2v}	b_{1v}	b_{0v}	c_{1v}	c_{0v}
1	1.0	- 0.753724	1.0	- 0.898974	0.954300
2	1.0	- 0.677901	1.0	- 0.867382	0.847751
3	1.0	- 0.473608	1.0	- 0.833090	0.691073
4	1.0	0.010521	1.0	- 0.790159	0.470015
5	1.0	1.049647	1.0	- 0.746811	0.238230
6	1.0	1.0		- 0.363595	
K = 0.005855		$q_{b1} = 5, q_{c1} = q_{co} = 6$			$w_c = 7$
v	b_{2v}	b_{1v}	b_{0v}	c_{1v}	c_{0v}
1	1.0	- 0.750000	1.0	- 0.906200	0.953125
2	1.0	- 0.687500	1.0	- 0.875000	0.843750
3	1.0	- 0.468750	1.0	- 0.828125	0.687500
4	1.0	- 0.000000	1.0	- 0.796875	0.468750
5	1.0	1.062500	1.0	- 0.750000	0.234375
6	1.0	1.0		- 0.359375	

1.1.7 erweiterter Filtergrad, Koeffizienten optimiert

7	n = 12	$\Delta n/n_{\min} = 0.5$, $C = 2.73$, $c = 0.166$			
K = 0.00477					$w_c = \infty$
v	b_{2v}	b_{1v}	b_{0v}	c_{1v}	c_{0v}
1	1.0	- 0.755086	1.0	- 0.891831	0.957336
2	1.0	- 0.692670	1.0	- 0.858725	0.860438
3	1.0	- 0.532074	1.0	- 0.820142	0.723601
4	1.0	- 0.174923	1.0	- 0.769618	0.529699
5	1.0	+ 0.583998	1.0	- 0.711996	0.302535
6	1.0	1.738215	1.0	- 0.669714	0.134694
K = 0.00469		$q_{b1} = 5$, $q_{c1} = q_{co} = 6$			$w_c = 7$
v	b_{2v}	b_{1v}	b_{0v}	c_{1v}	c_{0v}
1	1.0	- 0.75000	1.0	- 0.890625	0.953125
2	1.0	- 0.68750	1.0	- 0.859375	0.859375
3	1.0	- 0.53125	1.0	- 0.812500	0.718750
4	1.0	- 0.18750	1.0	- 0.765625	0.531250
5	1.0	- 0.59375	1.0	- 0.718750	0.296875
6	1.0	1.75000	1.0	- 0.671875	0.140625

1.2 Tschebyscheff-Verhalten im Durchlaßbereich

1.2.1 minimaler Filtergrad, gleichmäßige Ausnutzung bei ungerundeten Koeffizienten

8	n = 17	$\Delta n/n_{\min} = 0.032$, $C = 0.182$, $c = 0.6325$			
K = 0.929 E-9					$w_c = \infty$
v	b_{2v}	b_{1v}	b_{0v}	c_{1v}	c_{0v}
1	1.0	2.0	1.0	- 0.980243	0.977640
2	1.0	2.0	1.0	- 1.007306	0.934052
3	1.0	2.0	1.0	- 1.078352	0.891543
4	1.0	2.0	1.0	- 1.185395	0.850126
5	1.0	2.0	1.0	- 1.316566	0.810658
6	1.0	2.0	1.0	- 1.454961	0.775009
7	1.0	2.0	1.0	- 1.579033	0.745981
8	1.0	2.0	1.0	- 1.665826	0.726780
9	1.0	1.0		- 0.848550	
notwendige Wortlänge $w = 14$, $q_{c1} = 12$ $q_{co} = 13$					

1.2.2 minimaler Filtergrad, Koeffizienten optimiert

9	n = 17	$\Delta n/n_{\min} = 0.032 \quad C = 0.152, c = 0.0197$			
$w_c = \infty$			$w_c = 12, q_{c1} = 10, q_{co} = 11$		
v	c_{1v}	c_{ov}	v	c_{1v}	c_{ov}
1	- 0.977125	0.976002	1	- 0.977539	0.976074
2	- 1.002522	0.929336	2	- 1.002929	0.929199
3	- 1.071613	0.883973	3	- 1.071289	0.883789
4	- 1.176270	0.839916	4	- 1.176757	0.839843
5	- 1.304687	0.798061	5	- 1.304687	0.797851
6	- 1.439916	0.760363	6	- 1.439451	0.760253
7	- 1.561076	0.729744	7	- 1.561523	0.729980
8	- 1.645741	0.709528	8	- 1.645507	0.709472
9	- 0.838110		9	- 0.837890	
$b_{2v} = b_{ov} = 1, b_{1v} = 2, v = 1(1)8; b_{29} = b_{19} = 1,$					
K = 0.1048 E-8			K = 0.1044 E-8		

1.2.3 erweiterter Filtergrad, Koeffizienten optimiert

10	n = 18	$\Delta n/n_{\min} = 0.093, C = 0.0911, c = 0.0625$			
$w_c = \infty$			$w_c = 11, q_{c1} = 9, q_{co} = 10$		
v	c_{1v}	c_{ov}	v	c_{1v}	c_{ov}
1	- 0.970907	0.974429	1	- 0.970703	0.974609
2	- 0.989566	0.924861	2	- 0.990234	0.924804
3	- 1.047271	0.876898	3	- 1.046875	0.876953
4	- 1.137231	0.830441	4	- 1.136718	0.830078
5	- 1.250005	0.786144	5	- 1.250000	0.786132
6	- 1.372671	0.745559	6	- 1.373046	0.745117
7	- 1.488851	0.711109	7	- 1.488281	0.710937
8	- 1.580401	0.685738	8	- 1.580078	0.685546
9	- 1.631064	0.672205	9	- 1.630859	0.671875
$b_{2v} = b_{ov} = 1, b_{1v} = 2, v = 1(1)9$					
K = 0.372 E-9			K = 0.369 E-9		

1.3 Tschebyscheff-Verhalten im Sperrbereich
 minimaler Filtergrad, gleichmäßige Ausnutzung
 bei ungerundeten Koeffizienten

11	n = 17 $\Delta n/n_{\min} = 0.032$ C = 3247.12, c = 0.6322				
	K = 0.00713				$w_c = \infty$
v	b_{2v}	b_{1v}	b_{0v}	c_{1v}	c_{0v}
1	1.0	- 0.754196	1.0	- 0.891161	0.931981
2	1.0	- 0.693993	1.0	- 0.791694	0.803822
3	1.0	- 0.564634	1.0	- 0.657376	0.677838
4	1.0	- 0.347095	1.0	- 0.488242	0.547766
5	1.0	- 0.011940	1.0	- 0.288530	0.411972
6	1.0	0.472863	1.0	- 0.071949	0.275340
7	1.0	1.098435	1.0	0.133215	0.151513
8	1.0	1.719942	1.0	0.284959	0.062139
9	1.0	1.00		0.170776	
notwendige Wortlänge w = 11, $q_{b1} = 9$, $q_{c1} = q_{co} = 10$					

2. Bandpässe

Toleranzschema: $\Omega_{D1} = 0.87$, $\Omega_{D2} = 1.22$ $\Omega_{S1} = 0.75$, $\Omega_{S2} = 1.38$
 $\delta_D = 0.02$ $\delta_S = 0.01$

2.1 Approximation nach CAUER $n_{\min} = 7.974$

2.1.1 minimaler Filtergrad, Koeffizienten optimiert

12	n = 8 $\Delta n/n_{\min} = 0.003$ C = 4.465, c = 0.126				
	K = 0.0132				$w_c = \infty$
v	b_{2v}	b_{1v}	b_{0v}	c_{1v}	c_{0v}
1	1.0	- 0.336628	1.0	- 0.632809	0.931832
2	1.0	+ 0.498468	1.0	- 0.762763	0.800909
3	1.0	- 1.759163	1.0	- 1.061215	0.819601
4	1.0	- 1.479006	1.0	- 1.268594	0.944940
K = 0.0132 $q_{b1} = q_{c1} = 13$ $q_{co} = 14$ $w_c = 15$					
v	b_{2v}	b_{1v}	b_{0v}	c_{1v}	c_{0v}
1	1.0	- 0.336669	1.0	- 0.632812	0.931823
2	1.0	- 0.498413	1.0	- 0.762817	0.800903
3	1.0	- 1.759155	1.0	- 1.061157	0.819641
4	1.0	- 1.479003	1.0	- 1.268554	0.944946

2.1.2 erweiterter Filtergrad, Koeffizienten optimiert

13	n = 10	$\Delta n/n_{\min} = 0.157 \quad C = 2.36, \quad c = 0.165$			
	K = 0.00906				$w_c = \infty$
v	b_{2v}	b_{1v}	b_{0v}	c_{1v}	c_{0v}
1	1.0	- 0.354229	1.0	- 0.585997	0.930378
2	1.0	0.020479	1.0	- 0.636535	0.778289
3	1.0	0.000000	- 1.0	- 0.862342	0.698450
4	1.0	- 1.621733	1.0	- 1.145522	0.812458
5	1.0	- 1.470730	1.0	- 1.298018	0.945162
	K = 0.00896		$q_{b1} = q_{c1} = 7, \quad q_{co} = 8$		$w_c = 9$
v	b_{2v}	b_{1v}	b_{0v}	c_{1v}	c_{0v}
1	1.0	- 0.351562	1.0	- 0.585937	0.929687
2	1.0	0.023437	1.0	- 0.632812	0.777343
3	1.0	0.000000	- 1.0	- 0.859375	0.699218
4	1.0	- 1.625000	1.0	- 1.148437	0.812500
5	1.0	- 1.468750	1.0	- 1.296875	0.945312

3. Bandsperren

Toleranzschema entsprechend [54]

$$\Omega_{D1} = 1.626 \quad \Omega_{D2} = 1.767 \quad \Omega_{S1} = 1.631 \quad \Omega_{S2} = 1.782$$

$$\delta_D = 0.0559$$

$$\delta_S = 0.00018$$

3.1 Approximation nach CAUER $n_{\min} = 21.66$

3.1.1 minimaler Filtergrad, vollständige Ausnutzung des Durchlaßbereichs

14	n = 22	$\Delta n/n_{\min} = 0.0157 \quad C = 48,19, \quad c \approx 0.99$			
	K = 0.6148				$w_c = \infty$
v	b_{2v}	b_{1v}	b_{0v}	c_{1v}	c_{0v}
1	1.0	0.418608	1.0	0.428270	0.998906
2	1.0	0.414476	1.0	0.431500	0.995722
3	1.0	0.403398	1.0	0.441232	0.988277
4	1.0	0.379266	1.0	0.466045	0.966106
5	1.0	0.334773	1.0	0.518105	0.874176
6	1.0	0.271336	1.0	0.207472	0.529263
7	1.0	0.207337	1.0	-0.020204	0.869414
8	1.0	0.161771	1.0	0.061653	0.965146
9	1.0	0.136817	1.0	0.094057	0.987992
10	1.0	0.125306	1.0	0.106345	0.995624
11	1.0	0.121003	1.0	0.110615	0.998882

notwendige Wortlänge $w = 20 \quad q_{b1} = q_{c1} = q_{co} = 19$

3.1.2 minimaler Filtergrad, Koeffizienten optimiert

15	n = 22	$\Delta n/n_{\min} = 0.0157, \quad C = 44.32, \quad c = 0.501$			
K = 0.6191					$w_c = \infty$
v	b_{2v}	b_{1v}	b_{0v}	c_{1v}	c_{0v}
1	1.0	0.418608	1.0	0.428202	0.998862
2	1.0	0.414476	1.0	0.431350	0.995551
3	1.0	0.403398	1.0	0.440833	0.987824
4	1.0	0.379266	1.0	0.464775	0.964923
5	1.0	0.334773	1.0	0.512409	0.871679
6	1.0	0.271336	1.0	0.209672	0.545478
7	1.0	0.207337	1.0	-0.014777	0.866925
8	1.0	0.161771	1.0	0.062662	0.963935
9	1.0	0.136817	1.0	0.094347	0.987529
10	1.0	0.125306	1.0	0.106454	0.995450
11	1.0	0.121003	1.0	0.110674	0.998837
K = 0.6173		$q_{b1} = q_{c1} = q_{co} = 15$			$w_c = 16$
v	b_{2v}	b_{1v}	b_{0v}	c_{1v}	c_{0v}
1	1.0	0.418579	1.0	0.428192	0.998870
2	1.0	0.414489	1.0	0.431335	0.995544
3	1.0	0.403381	1.0	0.440826	0.987823
4	1.0	0.379272	1.0	0.464782	0.964935
5	1.0	0.334777	1.0	0.512420	0.871673
6	1.0	0.271362	1.0	0.209686	0.545471
7	1.0	0.207336	1.0	-0.014770	0.866912
8	1.0	0.161743	1.0	0.062652	0.963928
9	1.0	0.136840	1.0	0.094360	0.987518
10	1.0	0.125305	1.0	0.106445	0.995452
11	1.0	0.121032	1.0	0.110687	0.998840

3.1.3 erweiterter Filtergrad, Koeffizienten optimiert

16	n = 24	$\Delta n/n_{\min} = 0.108 \quad C = 33,70, \quad c = 0.271$			
K = 0.618					$w_c = \infty$
v	b_{2v}	b_{1v}	b_{0v}	c_{1v}	c_{0v}
1	1.0	0.418681	1.0	0.427641	0.998702
2	1.0	0.415284	1.0	0.429684	0.995141
3	1.0	0.406531	1.0	0.435648	0.987740
4	1.0	0.388125	1.0	0.449022	0.969118
5	1.0	0.354113	1.0	0.471886	0.913082
6	1.0	0.301928	1.0	0.417367	0.727137
7	1.0	0.240613	1.0	0.045555	0.720054
8	1.0	0.187600	1.0	0.040425	0.910454
9	1.0	0.152630	1.0	0.080472	0.968322
10	1.0	0.133566	1.0	0.099761	0.987452
11	1.0	0.124465	1.0	0.108081	0.995032
12	1.0	0.120926	1.0	0.111215	0.998673

K = 0.6163		$q_{b1} = q_{c1} = q_{co} = 13$			$w_c = 14$
v	b_{2v}	b_{1v}	b_{ov}	c_{1v}	c_{ov}
1	1.0	0.418701	1.0	0.427612	0.998657
2	1.0	0.415283	1.0	0.429687	0.995117
3	1.0	0.406494	1.0	0.435668	0.987792
4	1.0	0.388183	1.0	0.448974	0.969116
5	1.0	0.354003	1.0	0.471923	0.913085
6	1.0	0.302001	1.0	0.417358	0.727172
7	1.0	0.240722	1.0	0.045532	0.720092
8	1.0	0.187500	1.0	0.040405	0.910400
9	1.0	0.152587	1.0	0.080444	0.968261
10	1.0	0.133544	1.0	0.099731	0.987426
11	1.0	0.124511	1.0	0.108032	0.994995
12	1.0	0.120849	1.0	0.111206	0.998657

3.1.4 erweiterter Filtergrad, Pseudo-Gleitkomma- und Differenzdarstellung der Koeffizienten, optimiert

$$c_o = 1 - c_o', \quad e \geq -4$$

17	n = 24	$\Delta n/n_{\min} = 0.108 \quad C = 35.78, \quad c = 0.321$			
K = 0.615					$w_c = \infty$
v	b_{2v}	b_{1v}	b_{ov}	c_{1v}	c_{ov}
1	1.0	0.418681	1.0	0.427696	0.998726
2	1.0	0.415284	1.0	0.429792	0.995231
3	1.0	0.406531	1.0	0.435900	0.987960
4	1.0	0.388125	1.0	0.449689	0.969613
5	1.0	0.354113	1.0	0.473982	0.913971
6	1.0	0.301928	1.0	0.422542	0.724503
7	1.0	0.240613	1.0	0.039330	0.717130
8	1.0	0.187600	1.0	0.038448	0.911345
9	1.0	0.152630	1.0	0.079909	0.968827
10	1.0	0.133566	1.0	0.099558	0.987677
11	1.0	0.124465	1.0	0.107994	0.995124
12	1.0	0.120926	1.0	0.111164	0.998698
K = 0.6089		Pseudo-Gleitkomma-Darstellung			$i_M = 10$
v	b_{2v}	b_{1v}	b_{ov}	c_{1v}	c_{ov}
1	1.0	0.836914 -1	1.0	0.855468 -1	0.020508 -4
2	1.0	0.831054 -1	1.0	0.859375 -1	0.076172 -4
3	1.0	0.813476 -1	1.0	0.872070 -1	0.182383 -4
4	1.0	0.776367 -1	1.0	0.899414 -1	0.486328 -4
5	1.0	0.708007 -1	1.0	0.948242 -1	0.688477 -3
6	1.0	0.603515 -1	1.0	0.844727 -1	0.550781 -1
7	1.0	0.962896 -2	1.0	0.624491 -4	0.565430 -1
8	1.0	0.750000 -2	1.0	0.615234 -4	0.708984 -3
9	1.0	0.610351 -2	1.0	0.639648 -3	0.499023 -4
10	1.0	0.534179 -2	1.0	0.796875 -3	0.197266 -4
11	1.0	0.996093 -3	1.0	0.864258 -3	0.078125 -4
12	1.0	0.967773 -3	1.0	0.889648 -3	0.020508 -4

LEBENS LAUF

Günter F. Dehner

20.3.1944	Geburt in Würzburg
1950 - 1955	Grundschule in Ochsenfurt / Main
1955 - 1964	Röntgen - Oberrealschule Würzburg
1964	Abitur
1964	1. Teil Vorpraktikum bei Fa. Koenig & Bauer Würzburg
1964 - 1971	Studium der Elektrotechnik an der Technischen Universität München
1965	2. Teil Vorpraktikum bei Fa. Koenig & Bauer Würzburg
1967	1. Teil Hauptpraktikum bei Fa. Siemens München
1968	2. Teil Hauptpraktikum bei Fa. Rohde & Schwarz München
1968	Werkstudent bei Fa. AEG-Telefunken Ulm
1971	Abschluß des Studiums mit Diplomarbeit am Institut für Nachrichtentechnik
1971	wissenschaftliche Hilfskraft am Institut für Nachrichtentechnik der Tech. Univer- sität München
seit 1972	wissenschaftlicher Assistent am Institut für Nachrichtentechnik der Universität Erlangen-Nürnberg

G. F. Dehner

Copy-Center 2000

852 ERLANGEN · Obere Karlstr. 30
Tel. 09131-26197